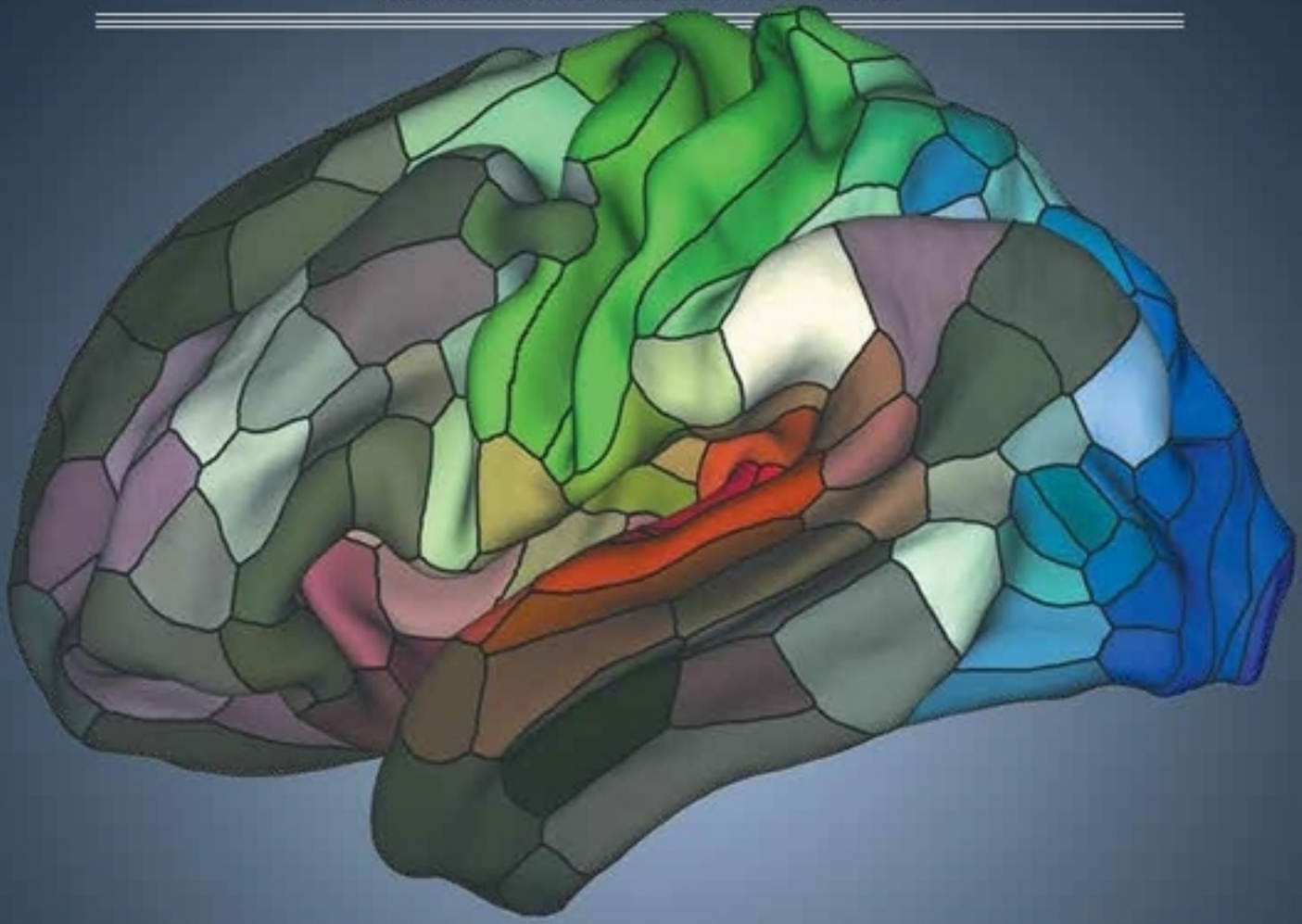
nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE



THE BRAIN REDEFINED

An updated map of the human cerebral cortex identifies 180 distinct brain regions per hemisphere PAGES 152 & 171

INFORMATION TECHNOLOGY

GIVE US MORE BANDWIDTH!

The battle to keep the Internet up to speed

PAGE 139

CONSERVATION

BIODIVERSITY'S OLD ENEMIES

Overexploitation and agriculture still main threats

PAGE 143

PLANETARY SCIENCE

JUPITER'S HOTTEST SPOT

Energetic waves from beneath the Great Red Spot

PAGE 190

NATURE.COM/NATURE

11 August 2016 £10

Vol 536, No. 7615

THIS WEEK

Energetic concerns

Rewarding existing nuclear power plants for the value of their low-carbon power makes sense, but the nuclear industry has a lot of work to do if it is survive and thrive in the twenty-first century.

When the state of New York moved last December to require utility companies to provide 50% of their power through renewable sources by 2030, questions about nuclear power naturally arose. Six nuclear reactors at four facilities currently provide more than 30% of the state's electricity — and more than half of its low-carbon source. Four of those plants were at risk of closure owing to simple economics: they have not been able to compete with cheap natural gas.

After factoring in the climatic value of low-carbon power generated at these stations, however, state regulators created a new subsidy on 1 August. The state began with the 'social cost of carbon', which represents the damage caused by greenhouse-gas emissions. The US government's central estimate is currently US$38 per tonne of carbon dioxide, rising to $50 in 2030. Revenues were well below that, so these plants will now be eligible for a 'zero-emissions credit' designed to make up the shortfall. In the first 2 years alone, that subsidy could be roughly $965 million. Illinois-based Exelon Corporation, which owns two of the facilities and is in negotiations to purchase the third, said it would press forward with its plan to keep the plants running.

The first lesson is that the price of carbon matters. New York is one of nine eastern states participating in an emissions trading system. The current price — averaging around $4 per tonne of CO_2 — was not high enough to keep nuclear power competitive with natural gas.

The US nuclear industry, and some pro-nuclear environmentalists, have hailed the New York standard as a precedent, and rightly so. It's a potential model for other US states in which nuclear power is facing similar economic hurdles. More generally, it's yet another reminder that climate policies have a long way to go, despite the rhetoric enshrined in the Paris climate agreement last year.

The nuclear industry's woes don't end there, however. Roughly 440 nuclear power plants currently provide 11% of the world's electricity, but they are on average 30 years old. More than 60 reactors are under construction, but the industry must work just to maintain its share of the energy mix as older plants close in the coming decades.

Simultaneously, New York state is opposing efforts to extend the lives of two other reactors at the Indian Point Energy Center on safety grounds. The operator has been fending off questions about tritium contamination in groundwater and various equipment malfunctions while applying for a permit from the US Nuclear Regulatory Commission to extend the life of the reactors from 40 to 60 years.

As long as nuclear power plants can demonstrate that they can operate safely, their contribution to the global effort to reduce greenhouse gases should be encouraged. But the reality is that there may be places where governments — and communities — decide that the potential price of a nuclear accident is too high. Whether the industry can expand in any meaningful way may depend on a new — and as yet unproven — generation of accident-proof reactors. Despite its efforts to keep a few reactors alive for now, New York is clearly betting on renewables. ■

CERN's road bump

The disappearing LHC signal is disappointing for those pitching for the next big accelerator.

Science thrives on discovery, so it's natural for physicists to mourn this week. As the high-energy-physics community gathered in Chicago on Friday, hopes were high (if cautious) that the Large Hadron Collider (LHC) at CERN, Europe's particle-physics laboratory near Geneva, Switzerland, had chalked up another finding to build on the discovery of the Higgs boson. Not so — the bump in the data that had caused such excitement was washed away with a flood of data that revealed it to be a mere statistical fluctuation.

Ordinarily, physicists would be satisfied if the LHC continued its bread-and-butter existence of confirming with ever-greater precision the standard model — a remarkably successful theory that is known to be incomplete. But the excitement over the bump has left them hungry for more. As is evident from the 500 theory papers written about the bump, physics is ready for something new.

That the LHC has not turned up anything beyond the standard model does not mean it never will. The machine has collected just one-tenth of the data that scientists hoped to amass by the end of 2022, and just 1% of those it could collect if a planned revamp to increase the intensity of collisions goes ahead. But the dry spell worries some. The idea of supersymmetry predicts that heavier counterparts to regular particles will become evident at higher collision energies. Before the LHC was switched on, fans of the theory would have gambled on being able to see something by now. And if the dry spell extends to a drought, high-energy physics could descend into what some call the nightmare scenario — the collider finds nothing beyond the Higgs boson. Without 'new' physics, there is no thread to pull to unravel the countless mysteries that the standard model fails to account for, including dark matter and gravity.

There remain strong reasons to build a successor machine. But without another discovery, the public's delight in high-energy physics could fade: there comes a time when exploration alone no longer satisfies.

Convincing funding agencies to cough up several billion dollars to continue the same approach will therefore be tough, especially when neutrino and lab-based precision experiments cost a fraction of the price. It will be physicists' job to consider carefully the worth of pursuing that discovery strategy. And if high-energy colliders remain essential, they need to work on their sales pitch. ■


JOHN BARUCH

Steer driverless cars towards full automation

For cars to be safe, full control must be allocated to the driver — be it human or computer, argues **John Baruch**.

As the popularity of self-driving cars increases, so do the concerns. Last month, China banned tests of autonomous vehicles on public roads. And investigations continue into the death of Joshua Brown, who was killed when his Tesla car on autopilot ploughed into the side of an articulated truck in Florida. His car used visible-light cameras to image the road and computers to evaluate the situation. But according to Tesla, the white truck merged with the bright Florida sky and was not recognized.

Self-driving vehicles promise to extend and enrich travel. But they raise profound questions for society about our relationship with machines. How will people cope, and what business models will develop?

Will car manufacturers emulate jet-engine manufacturers such as Rolls-Royce and Pratt & Whitney and start to sell travel distance and continuously record the operation of every engine they sell? Will car owners use the Uber model to rent out their vehicles as taxis instead of leaving them in a car park? Will automation revolutionize rural transport and give the rural poor, young, old and disabled the low-cost travel they are entitled to?

The challenges for the architects of our city centres, housing, streets, schools and workplaces are immense. And the issues around control of information such as big data and privacy must be confronted. Car manufacturers may find that the information they glean from tracking the lifestyle of their customers is worth much more than their vehicles.

The scientific community and society in general need to engage with these questions, and together decide what kind of future we want and how autonomous vehicles fit in. If we don't, the future is likely to be mapped out by companies that merely want to make money from the technology. The questions are complex, but they can be boiled down to one: should self-driving cars have a steering wheel?

The big Internet companies such as Google, Apple and Baidu — those generating the real pressure for self-driving vehicles — do not think that they should. These companies are keen to maximize time online for those who are rich enough to afford a car. Google's business model can use daily commute time that is no longer spent driving a car to increase the value of its advertising. It would therefore not want self-driving vehicles that allow the driver to take over. The vehicle is totally autonomous.

A number of driverless vehicles with no steering wheel and no opportunity for people to take control are on trial — including a parking transit system at London Heathrow Airport and buses in the Netherlands, Italy and China. It is the Chinese who are taking the opportunity most seriously, where Internet companies are working with vehicle manufacturers.

But this model poses a problem for Tesla and for many other car manufacturers, especially for the more expensive brands. The appeal to customers of luxury car brands tends to be the driving experience. And if the car has no steering wheel — and the 'driver' is a mere passenger — that appeal evaporates. That's why Tesla, Jaguar Land Rover and others use the technology in existing self-driving cars only to provide support, with the driver officially remaining in charge. Formally, Brown was in charge of the vehicle he died in.

Car and component companies are working hard to generate a business model for more-autonomous vehicles that have steering wheels. It is clear that they remain keen to provide the driver with assistance, and that means there is a real need for research: on both the technical and social-science aspects. How will drivers react when the car tells them to take over? How can the handover be made safe? How can the vehicle be brought to a safe halt if the driver does not take charge?

INCREMENTAL SUPPORT IS NOT A SAFE COMPROMISE.

I operate an autonomous robotic telescope in the Canary Islands that is 3,000 kilometres from its base in the United Kingdom. Autonomous telescopes do not pose the same dangers to the public as self-driving vehicles do, but there is a lot that can be learnt from our experiences. We have removed nearly all the single-point failure modes by quadruplexing all the crucial information flows (when one fails, you can still poll the others and isolate the failure) and have instituted an artificial-intelligence reconfiguration process that isolates failure until it is repaired. With quadruplexed systems, Brown might not have died. The car would have slowed down, if only because the sensor systems were experiencing confusion.

The driver-support philosophy is a flawed approach. Some aircraft already have technology that can take the plane from runway to runway, with the pilot just taxiing the plane to and from the stand. It is generally not used, because, with no role in the flight, pilots can become bored and do other things. They are then totally unprepared to take over if needed. Incremental support is not a safe compromise. People must either drive a car or be driven by it.

The efforts of the luxury car brands are reminiscent of when gas companies tried to improve lighting by adjusting lantern mantles when electric lighting appeared. For self-driving cars to rule the road, the steering wheel must go the way of the Model T Ford. ■

John Baruch *is a senior lecturer at the University of Bradford, UK, and visiting professor at South China University of Technology in Guangzhou.*
e-mail: j.e.f.baruch@bradford.ac.uk

RESEARCH HIGHLIGHTS

Selections from the scientific literature

GEOCHEMISTRY

Minerals mimic synthetic structure

Researchers have found naturally occurring metal–organic frameworks (MOFs) — chemical structures that were thought to exist only when made in the lab.

MOFs have open, porous architectures, which could make them useful in catalysis, photovoltaics and other applications. Tomislav Friščić at McGill University in Montreal, Canada, Sergey Krivovichev at Saint Petersburg State University in Russia and their colleagues used X-ray diffraction to study two samples from a permafrost drill core, which was taken from a Siberian coal mine 230 metres below Earth's surface more than 70 years ago. They observed that the rare organic minerals stepanovite and zhemchuzhnikovite contain channels, pores and other structures that are found in synthetic MOFs.

These are the only organic minerals known so far to have open architectures, the authors say.

Sci. Adv. 2, **e1600621 (2016)**

URBAN ECOLOGY

Insect mix high in rich areas

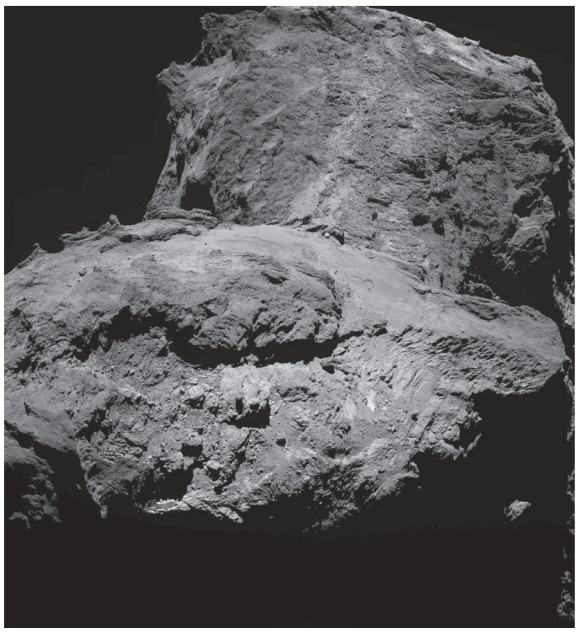
The interiors of homes in affluent neighbourhoods host a wider diversity of insects and spiders than do those in less wealthy areas.

Neighbourhoods with a high income often have a higher diversity of plants and certain animals, such as birds, than other areas. To find out whether this 'luxury effect' extends indoors, Misha Leong at the California Academy of Sciences in San Francisco and her colleagues sampled all arthropods — living and dead — including insects (**pictured** is *Sciara hemerobioides*), spiders and millipedes, inside 50 homes in and around Raleigh, North Carolina. They found that arthropod diversity increased with house size and diversity of surrounding vegetation, and were surprised to find a strong influence of average neighbourhood income, too.

Affluence could be affecting arthropod diversity through urban planning and landscaping at the neighbourhood level.

Biol. Lett. 12, **20160322 (2016)**

IAN REDDING/GETTY



ESA/ROSETTA/MPS FOR OSIRIS TEAM

ASTRONOMY

Gentle birth of a comet

The comet 67P/Churyumov–Gerasimenko (pictured), which has been orbited by the Rosetta spacecraft since 2014, might date back to the primordial Solar System billions of years ago.

A team led by Björn Davidsson at NASA's Jet Propulsion Laboratory in Pasadena, California, used instruments on the European Space Agency's spacecraft to examine the structure of the comet's core. The porous consistency of 67P shows that it did not form through violent collisions. Instead, the authors propose that the comet was made gradually, when icy pebbles from the outer reaches of the developing Solar System clumped together. The two lobes of 67P may have gently joined together during the final stages of the comet's formation.

Astron. Astrophys. 592, **A63 (2016)**

PALAEOECOLOGY

Thirst finished off the mammoths

One of the last woolly mammoth populations died out on an island off the coast of Alaska nearly 6,000 years ago, probably because of a shrinking supply of fresh water.

Human hunting has been linked to the extinction of the species (*Mammuthus primigenius*), but this relict population perished without our help, according to Russell Graham of Pennsylvania State University in University Park and his colleagues. The authors examined ancient DNA, isotopes and plant and animal material in sediment cores from a lake on St Paul Island. They also studied mammoth fossils. The researchers estimate that the island's mammoths became extinct 5,600 years ago, when the island was shrinking because of sea-level rise and the lake was evaporating into a salty puddle — perhaps because of long-standing drought, or depletion by the mammoths themselves.

Freshwater scarcity could drive island extinctions more often than previously thought, the authors say — and will only increase as the climate changes.

Proc. Natl Acad. Sci. USA **http://doi.org/bm9z (2016)**

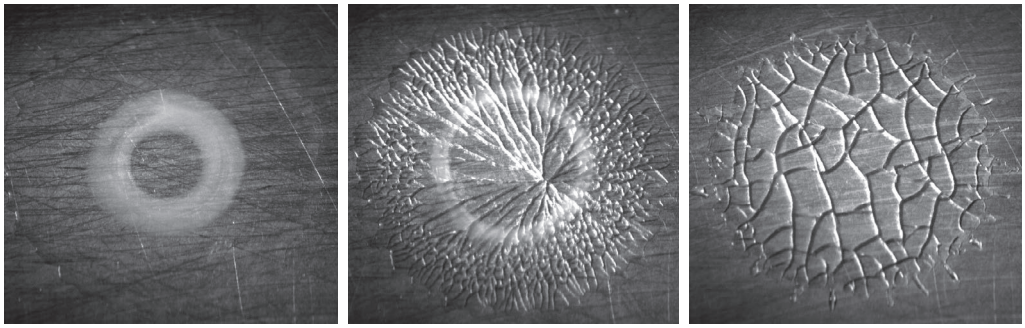
PARTICLE PHYSICS

No sign of new neutrino

A massive detector at the South Pole has found no evidence of a 'sterile' neutrino: a near-massless particle that is thought to interact only through gravity.

Hints of this possible fourth type of neutrino first emerged in the 1990s, and were rekindled early this year by an experiment in China. In the latest work, researchers

ARXIV.ORG



at the IceCube Neutrino Observatory in Antarctica, led by Francis Halzen at the University of Wisconsin–Madison, counted neutrinos of a known type that hit the detector from below. A dearth of these neutrinos at particular energies would have revealed that some of the particles had temporarily mutated into sterile neutrinos during their trip through Earth, but the researchers found no such feature in their data.

The experiment did not rule out the existence of heavier sterile neutrinos. A fourth kind of neutrino would challenge the standard model of particle physics, which allows for only three neutrino types.

Phys. Rev. Lett. 117, **071801 (2016)**

EVOLUTION

Ancient whales heard high notes

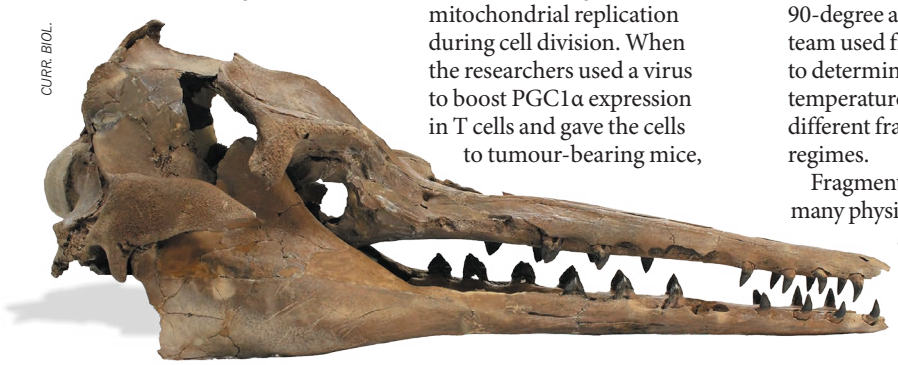
Fossil evidence suggests that the first whales could detect high-frequency sounds.

Researchers have debated whether animals called archaeocetes — the common ancestors of all modern whales and dolphins — specialized in hearing high frequencies, like modern killer whales, or low frequencies, like today's humpback whales. Morgan Churchill at the New York Institute of Technology in Old Westbury and his colleagues describe a new species of whale (fossil skull **pictured**) dating from 27 million to 24 million years ago. Features of its remarkably well-preserved inner ear, as well as other structures, suggest that the animal could generate and hear high-frequency sounds. The inner ear also has primitive features similar to those of archaeocetes.

The authors suggest that the first whales could hear higher frequencies than their terrestrial ancestors — an ability co-opted by later toothed whales for echolocation.

Curr. Biol. **http://doi.org/bnh5 (2016)**

CURR. BIOL.



CANCER IMMUNOLOGY

Immune cells tire out in tumours

After they invade tumours, immune cells gradually lose their ability to produce energy.

Greg Delgoffe and his colleagues at the University of Pittsburgh in Pennsylvania studied immune cells called T cells in mice with implanted tumours. They found that T cells inside tumours were less effective at taking up glucose than those in other parts of the body. The tumour-infiltrating cells also showed reduced total mass of mitochondria — cell organelles that produce energy — and contained abnormally shaped mitochondria. The metabolic defects were linked to reduced levels of PGC1α, a protein that regulates mitochondrial replication during cell division. When the researchers used a virus to boost PGC1α expression in T cells and gave the cells to tumour-bearing mice, the tumours shrank more and the animals lived longer than those that received non-reprogrammed cells.

Boosting metabolic processes in immune cells could help to improve cancer therapies, the authors say.

Immunity **http://doi.org/bndn (2016)**

PHYSICS

Crack patterns in freezing water

Water droplets landing on a cold surface fragment into one of two different patterns as they freeze, depending on the temperature of the surface.

Elisabeth Ghabache and her colleagues at the University of Pierre and Marie Curie in Paris used a high-speed camera to monitor the behaviour of pancake-shaped water droplets that froze on a cold steel surface after being dropped from a height of 36 centimetres. They observed no crack formation when the surface was at –20 °C (**pictured** left). At –30 °C and –40 °C, cracks spread from a central point towards the 'pancake' edge (centre). At –50 °C and –60 °C, the cracking occurred in a step-by-step manner, with the initial cracks splitting into newer ones at roughly 90-degree angles (right). The team used fracture modelling to determine the transition temperatures between the different fragmentation regimes.

Fragmentation occurs in many physical processes, such as bubble bursting and glass breaking. This model system could help researchers to learn more about various fracture mechanisms, the authors say.

Phys. Rev. Lett. **http://dx.doi.org/10.1103/physrevlett.117.074501 (2016)**

MICROBIOLOGY

Toxic bacteria adapt fast

Harmful blue-green algae can adapt rapidly to changing environments.

The photosynthetic cyanobacterium *Microcystis* produces toxic blooms in lakes and reservoirs. To test how different strains respond to changing carbon dioxide levels in water, Jef Huisman and his colleagues at the University of Amsterdam kept mixed populations in the laboratory and aerated the water with bubbles containing low or elevated levels of CO_2. In low CO_2 conditions, strains whose carbon-uptake systems are efficient when carbon is limited became dominant. When CO_2 was elevated, however, strains that have systems with high uptake rates outcompeted the others. The team studied *Microcystis* collected from Lake Kennemermeer in the Netherlands and found that the abundance of each strain shifted with seasonal changes in CO_2 availability.

Cyanobacteria may be more adept at dealing with high CO_2 levels than previously thought.

Proc. Natl Acad. Sci. USA **http://doi.org/bnf9 (2016)**

SEVEN DAYS *The news in brief*

PEOPLE

Femtochemist dies

Ahmed Zewail, the winner of the 1999 Nobel Prize in Chemistry, died on 2 August, aged 70. He is credited with founding the field of femtochemistry, which probes the mechanics of chemical reactions using laser pulses lasting just tens of femtoseconds (1 femtosecond is 10^{-15} s). An Egyptian-born US citizen working at the California Institute of Technology in Pasadena, Zewail was the first Arab to win a science Nobel. He championed science education and research in his native country and founded the Zewail City of Science and Technology, a university that opened its doors to students in 2012 near Cairo.

RESEARCH

Cancer drug fails

A promising lung-cancer drug failed in a key clinical trial, sending stock in the drug's developer, Bristol-Myers Squibb, tumbling by 17% on 5 August. The drug, called Opdivo (nivolumab), is one of a suite of new cancer therapies that release immune responses against tumours by blocking a protein called PD-1. Opdivo has been shown to benefit some people with advanced cancers, including lung cancer. But Bristol-Myers Squibb, which is based in New York City, announced last week that Opdivo had failed as a front-line therapy for lung cancer. Stock in a competing firm, Merck of Kenilworth, New Jersey, rose 10% after the news.

Turkish purge

The Turkish research agency TÜBİTAK in Ankara has removed 139 staff from their posts, pending investigations into their possible connections with the Gülen movement, which President Recep Tayyip Erdoğan claims was behind the country's attempted coup last month. A further 28 staff have resigned, according to statements by Turkish science minister Faruk Özlü on 5 August. TÜBİTAK, which helps to design the country's research policy and distributes grant money, has previously been purged. In 2014, agency engineers were dismissed after they declared that incriminating recordings of allegedly tapped telephone conversations between Erdoğan and his son — which Erdoğan said were fabricated — were not manipulated.

RAHEB HOMAVANDI/REUTERS

Iran executes scientist

Shahram Amiri, an Iranian nuclear scientist, has been hanged for espionage, his country's officials announced on 7 August. Amiri said that he had been abducted by the CIA during a pilgrimage to Saudi Arabia in 2009, and taken to the United States to be interrogated and tortured. But the US government denied that, and US media alleged that he had been a paid informant who defected voluntarily. He returned to Iran in 2010 and was later convicted of passing secrets about Iran's nuclear activities to the United States. Iran has maintained that its nuclear programme has only peaceful purposes. Amiri was reportedly an isotope researcher at the Malek-Ashtar University of Technology in Tehran.

Mosquito trial

The US Food and Drug Administration announced on 5 August that a proposed Florida field test of genetically modified mosquitoes poses few risks to the environment or human health. The announcement clears the way for the board of the Florida Keys Mosquito Control District to allow the release of *Aedes aegypti* mosquitoes carrying a gene that kills the insects' offspring, in an effort to control diseases they transmit, including dengue and Zika viruses. Local residents will vote in a non-binding referendum, currently slated for November, before the board decides whether to go ahead with the trial.

EVENTS

Missing vaccines

An investigation by the Associated Press has reported that 1 million yellow-fever vaccines sent in February by the World Health Organization and its partners to tackle a large outbreak in Angola cannot be accounted for. Six million vaccines were sent in total. Of those that can be traced, some were sent to regions where there was no yellow fever; others were improperly stored, or arrived without the syringes to administer them, according to the 5 August report. The agencies involved have responded that a wastage of around 10% is expected in mass-vaccination campaigns for yellow fever — and Angolan officials have denied that any vaccines went missing.

Goodnight Yutu

China's moon rover Yutu, or Jade Rabbit, was officially declared dead by state officials on 3 August. It arrived on the Moon in December 2013, and was intended to carry out a three-month exploration of the lunar surface, but it survived for more than two years before going dark for the last time. The six-wheeled, solar-powered rover was struck by mechanical difficulties in early 2014, but had already used its penetrating radar to probe the structure of the lunar soil to a depth of more than 100 metres, and sent back data and high-resolution images to Earth. The mission made China the third country

to land a craft on the Moon, after the United States and the Soviet Union.

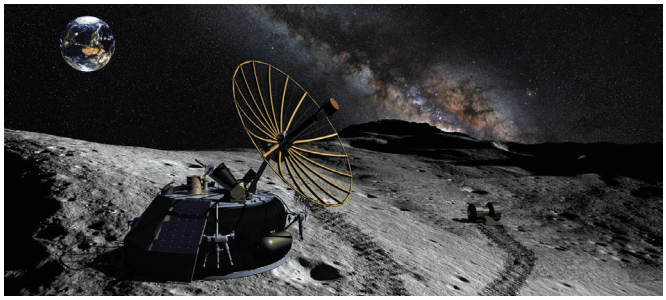
APS moves meeting

A division of the American Physical Society (APS) has voted to move its 2018 annual meeting from its originally planned location of Charlotte, North Carolina, as a result of a state law, enacted in March, that forces transgender people to use only toilets that correspond to their sex at birth. The chair of the APS Division of Atomic, Molecular, and Optical Physics said that it wanted to "provide a welcoming environment for all members". The APS, which made the statement on 4 August, will hold its conference in Florida instead.

Private Moonshot

A private mission to the Moon has been approved for the first time, by the US government. Moon Express, a company in Cape Canaveral, Florida, announced on 3 August that it had been given permission to travel beyond Earth's orbit and put a robotic lander on the Moon in 2017. Moon Express was founded in 2010 by three technology entrepreneurs, and is one of the companies competing for the Google Lunar XPRIZE; the competition will award US$20 million to the first company to land a privately funded spacecraft on the Moon. Moon Express is yet to finish building its MX-1 lander (**pictured**, artist's impression).

COURTESY OF MOON EXPRESS



Zika in Cuba

Cuba has discovered two cases of locally acquired Zika-virus infection, the nation's health ministry announced on 4 August. The country's preventive campaigns had been largely successful in staving off infections, with only 30 imported cases identified in Cuba this year, and only one previous locally transmitted case, in March. On 3 August, the US National Institutes of Health (NIH) launched the first clinical trial of a Zika vaccine, which it plans to test in 80 healthy volunteers. NIH officials say that the vaccine will probably not be ready for deployment until 2018.

POLICY

Data push-back

A coalition of researchers has rebuffed a proposal to share clinical-trial data rapidly. In February, the International Committee of Medical Journal Editors (ICMJE) proposed that clinical-trial leaders should make the de-identified patient data that underlie a journal article public within six months of publication. The 282 signatories to an article published on 4 August in *The New England Journal of Medicine* said that the ICMJE proposal was too burdensome and would have unintended consequences, such as delaying publication of results (*N. Engl. J. Med.* **375,** 405–407; 2016). They said that researchers should have at least two years — but up to five — to make data public.

US climate rule

The White House released a sweeping new climate policy on 2 August, instructing all federal agencies, from the Department of Agriculture to the Department of Transportation, to consider the impacts of their actions on climate change from now on. The agencies are also required to quantify those impacts, mainly in terms of greenhouse-gas emissions. The policy comes from the White House Council on Environmental Quality, which was established in 1969 to advise agencies when they are preparing environmental-impact statements.

FUNDING

Climate-cut U-turn

Australia's government has ordered its national science agency to re-prioritize basic climate research, six months after the organization unveiled controversial plans to slash jobs in the sector. The Commonwealth Scientific and Industrial Research Organisation (CSIRO) will — on government instructions — create 15 new climate-science jobs and receive an extra Aus$37 million (US$28 million) over the next 10 years, science minister Greg Hunt announced on 4 August. But the intervention may have come too late to repair damage already caused, researchers say. See go.nature.com/2akgeyp for more.

COMING UP

21–25 AUGUST
Scientists come together in Philadelphia, Pennsylvania, for the American Chemical Society's national meeting and exposition, where topics will include the science behind Pixar.
go.nature.com/2ayecj7

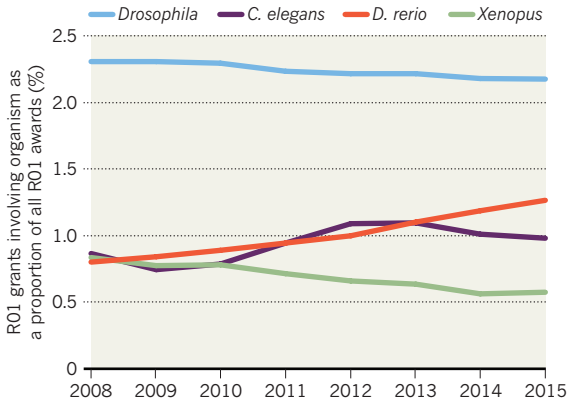
20–30 AUGUST
Kuala Lumpur hosts the 34th biennial conference of the Scientific Committee on Antarctic Research.
www.scar2016.com

TREND WATCH

Zebrafish (*Danio rerio*) are the rising stars of model-organism research, an analysis of grants from the US National Institutes of Health (NIH) finds. A team at the NIH Office of Portfolio Analysis used text mining and manual searching to assess successful applications for R01 awards, the NIH's main grant programme for individual investigators. Whereas the proportions of grants allocated to zebrafish and *Caenorhabditis elegans* studies rose between 2008 and 2015, the fraction for research with *Xenopus* frogs fell.

ZEBRAFISH COURT FUNDING DOLLARS

Grants for zebrafish (*Danio rerio*) research from the US National Institutes of Health's R01 award programme are on the rise.

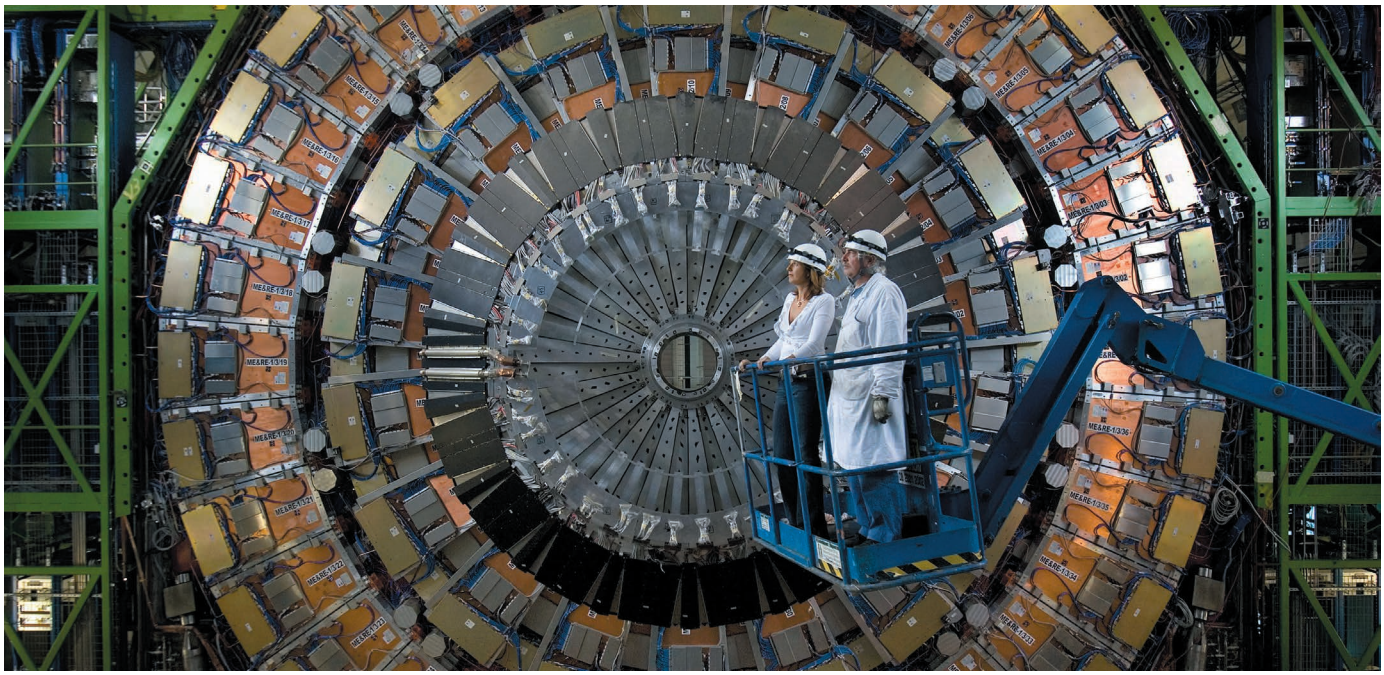


SOURCE: MICHAEL LAUER/NIH OFFICE OF PORTFOLIO ANALYSIS; GO.NATURE.COM/2ALCEOY

NATURE.COM
For daily news updates see:
www.nature.com/news

NEWS IN FOCUS

STEFANO DAL POZZOLO/CONTRASTO/EYEVINE



The CMS (pictured) was one of two experiments at the Large Hadron Collider that saw hints of an unexpected particle.

PARTICLE PHYSICS

LHC particle hopes dashed

Promising two-photon signal disappears as data pile up.

BY ELIZABETH GIBNEY, CHICAGO, ILLINOIS

It would have marked the beginning of a new era in particle physics. But the latest data have squashed hopes that hints of an unexpected particle detected by the Large Hadron Collider (LHC) would solidify with time. Instead, the intriguing data 'bump' first reported in December turns out to be just a statistical fluctuation.

Representatives from ATLAS and CMS — two independent experiments at the LHC, part of the European particle-physics laboratory, CERN — presented the news at the International Conference on High Energy Physics (ICHEP) in Chicago, Illinois, on 5 August. The analyses included nearly five times the amount of data used in December, and show that the signal has faded to almost nothing.

"There is no significant excess seen in the 2016 data," said Bruno Lenzi, an ATLAS physicist based at CERN near Geneva, Switzerland, to a standing-room-only session at ICHEP.

Additional data from CMS also failed to produce a significant signal, says Chiara Rovelli, a physicist at the National Institute of Nuclear Physics in Rome.

The announcement was a disappointment to researchers, but it wasn't unexpected. The ATLAS team's previous update, in June, put the signal's significance — a measure of the chances that random fluctuations in the data would produce such a bump without a particle — at 2.1 sigma. That was well below the 5-sigma threshold for determining whether a signal is a discovery or just noise.

But both ATLAS and CMS independently saw the signal, comprised of slightly more pairs of photons — with a combined energy of 750 gigaelectronvolts — than expected. That gave physicists hope that the bump was real. Researchers around the world produced more than 500 papers trying to explain the potential particle.

"Seeing a glimpse of something, even the half a glimpse that makes you hold your breath a moment and think, 'what if' — it's too valuable to be left unexplored," says Tara Shears, a particle physicist at the University of Liverpool, UK.

HISTORY OF A BUMP

The cautious excitement was driven by the bump's potential pay-off, says Don Lincoln, a physicist at the Fermi National Accelerator Laboratory near Batavia, Illinois. The standard model is incomplete because it fails to account for mysteries such as dark matter, and can't reconcile quantum mechanics with gravity. A new particle would have directed ▶

physicists towards an alternative theory, says Lincoln.

The signal was appealing in part because the analysis behind it was relatively simple and robust, says Christoffer Petersson, a theoretical physicist at Chalmers University of Technology in Gothenburg, Sweden.

The fact that the particle could have been a heavier cousin to the Higgs boson was also enticing, says Guido Tonelli, a physicist at the University of Pisa in Italy and former head of CMS.

Even though all those models are now wrong, it was a fun and useful exercise to try to explain the bump, says Petersson.

Statistical fluctuations and discoveries look identical at first, says Lincoln. Such coincidences are always possible when performing thousands of searches across a wide range of particle masses. It has happened before and will probably happen again, he says.

ONWARD

This false alarm does not affect the LHC's chances of finding something else, says Petersson. For now, it is business as usual for the collider's experiments.

Still, there is some concern that 40 years after the development of the standard model, particle accelerators, including the LHC, have not found anything beyond it.

It's surprising that nothing unexpected has emerged from the LHC data, says Guy Wilkinson, a physicist at the University of Oxford, UK. This underscores a growing unease in the community: as time goes on without new findings, it becomes less likely that the most appealing versions of supersymmetry — arguably the most promising way to extend the standard model — are true.

But Petersson notes that the chances that the LHC will find something beyond the standard model will go up this year and next, because the collider is operating near its maximum energy of 14 teraelectronvolts. If new particles are rare, or if they decay in ways that are hard to observe, they could take a while to emerge, he says.

And there are other ways of finding new particles, says Shears. With enough data, particles that are too heavy to be produced directly could reveal themselves through subtle influences on well-known particles. Physicists with LHCb, another experiment at the collider, have already found such hints, but they need more information to confirm them.

"We know already that sooner or later, one of these anomalies will survive all controls and suddenly — crack — everything will change," Tonelli says. "The beauty of our work is that this could happen at any time." ■ SEE EDITORIAL P.125

Perlan 2 aims to break the glider altitude record of 15,445 metres.

JAMES DARCY/AIRBUS GROUP

ATMOSPHERIC RESEARCH

Surfing glider set to study climate

Perlan mission will ride stratospheric waves and conduct atmospheric research.

BY DECLAN BUTLER

A glider that aims to soar higher than any other piloted aircraft will begin its first campaign this month in the skies above Argentina. For its pilots and engineers, the Perlan Project holds the excitement of breaking the world altitude record for gliding — and perhaps one day reaching close to the vacuum of space.

But for Elizabeth Austin, the project's chief scientist, there's another thrill: the glider will carry scientific instruments for climate, aerospace and stratospheric research that cannot be done using other means. "The possibilities are just so incredible," says Austin, an atmospheric physicist and the founder of forecasting service WeatherExtreme in Incline Village, Nevada.

The carbon-fibre glider, built with a pressurized cabin, is intended to achieve sustained flight at around 27,000 metres, where the density of air is about 2% of that at sea level. In the series of flights that the craft will begin in mid-August, it will fly to only 15,000–18,000 metres — in part because of weather conditions — but this could still break the glider altitude record of 15,445 metres, set by an earlier Perlan model.

The glider will carry instruments to measure levels of aerosols and greenhouse gases, including ozone, methane and water vapour, and will gather information on the exchange of gases and energy between the two lower layers of Earth's atmosphere: the troposphere and the stratosphere. Those data, to be collected this year and next, could improve climate models, which account poorly for these atmospheric interactions and contain "horrific" uncertainties about the levels and behaviour of water vapour at stratospheric altitudes, Austin says.

Scientific balloons have already flown at much higher altitudes, but they must follow the wind, Austin adds, whereas a pilot can steer and circle a glider. "We can spend hours flying where we want. A glider is an incredible scientific platform as there's no other way to get this sort of data."

"It's an extremely exciting project," says Jie Gong, an expert in atmospheric dynamics at NASA's sciences and exploration directorate in Greenbelt, Maryland. On the basis of its intended flight route, the Perlan glider might be able to provide the first direct observations of polar stratospheric clouds, a unique type of ice cloud that forms in the polar stratosphere and helps to deplete ozone, Gong adds.

The glider is named after those same clouds, which have an iridescent mother-of-pearl appearance (Perlan means 'pearl' in Icelandic). They are typically generated at high altitudes by stratospheric mountain waves — when strong winds that blow over the tops of high mountains are driven up towards space. In 1992, a retired NASA test pilot,

Einar Enevoldson, founded the Perlan Project with the aim of creating a glider that could surf these waves up to the stratosphere. And in 2006, he and the US adventurer Steve Fossett proved the concept with their record-breaking flight on Perlan 1, a modified conventional glider.

But Fossett's death the following year in a light-aircraft accident set the project back until July 2014, when European aerospace group Airbus became a major sponsor and contributed its research expertise. The Perlan 2 craft made its maiden flight last year in Oregon, and in March surfed its first mountain waves above the Sierra Nevada range in California.

Its next flights will be over El Calafate on the eastern and southern fringes of the Andes range in Argentina. There, during the South Pole's winter, a fast-moving, high-altitude jet stream called the polar-night jet extends from the troposphere into the upper atmospheric layers — helping the Andes mountain waves (and the glider) to reach the stratosphere (see 'Science on a glider').

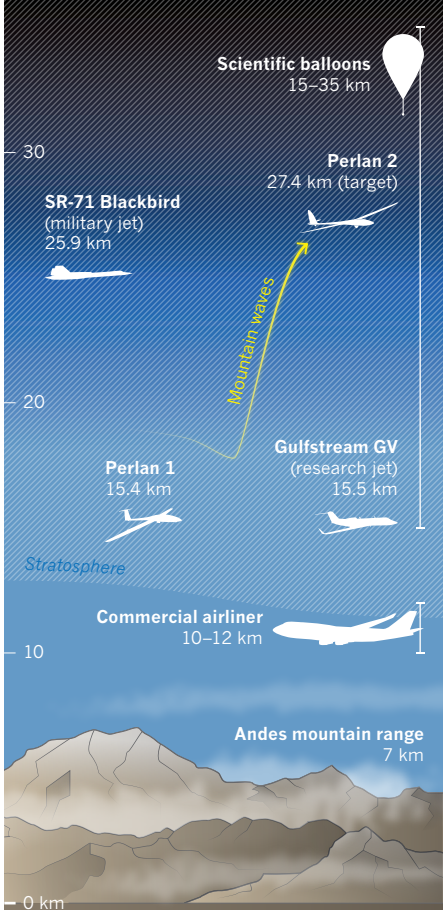
Besides its atmospheric chemistry, Perlan 2 will carry instruments to study turbulence in stratospheric mountain waves, and to explore the microphysics of interactions between mountain waves and polar meteorology, which ultimately affect weather variability. Information on how mountain waves break in the stratosphere is "extremely limited", says Gong, and requires detailed, fine-scale data on temperature, humidity and wind, which the glider is uniquely placed to measure. Airbus says that many of the weather phenomena Perlan 2 will encounter will provide useful information for it and other aircraft makers that are contemplating operating aeroplanes at higher altitudes.

Once Perlan is fully tested, says Austin, she hopes to get funding to use the glider as a long-term scientific platform that would examine how hourly, seasonal or even decadal changes in the stratosphere affect weather and climate.

A drone that could carry more instruments is a future possibility — but for now, a piloted craft is preferable and simpler, says Ed Warnock, the project's chief executive. Machines cannot yet match the best human pilots when it comes to climbing waves in such demanding flight conditions, he says.

Perlan's backers hope that it can surpass 27,000 metres in 2017 — and, ultimately, they intend another version of the glider to fly higher than 30,000 metres, where the air density is almost identical to that on Mars's surface. That might provide insight into how winged aircraft could fly on the red planet.

For now, engineers and scientists alike are just hoping to see the glider soar into the stratosphere above the Andes and take data. "Everything in the aircraft is experimental. It's a very difficult mission to do right, and to do it safely is not easy," Austin says. ■



SCIENCE ON A GLIDER
The Perlan glider aims to fly higher than any other piloted aircraft, and to conduct unique stratospheric research.

PAUL JACKMAN/*NATURE*

BIOMEDICINE

US to lift ban on funding for human–animal hybrids

Researchers in the United States will soon be able to resume chimaera-based projects.

BY SARA REARDON

Since September 2015, researchers have been banned from receiving funding from the US National Institutes of Health (NIH) for adding human stem cells to animal embryos, creating blends called chimaeras. But an NIH proposal released on 4 August lifts that moratorium, with certain exceptions. It also sets up a panel to review the ethics and oversight of grant applications.

The proposal shortens the window during which human cells can be introduced into non-human primate embryos, disallowing it before the central nervous system begins to form. This limits the number of human cells incorporated into a chimaera's brain. It also prohibits breeding animals containing human cells, preventing growth of a chimaeric embryo in a non-human womb or the birth of an animal more humanized than its parents. Grant applications that fall into a grey area would undergo a panel review.

The panel will pay particular attention to projects involving primates, mammals at very early developmental stages or those in which human cells could affect an animal's brain. Past a certain point, rodent embryos with human cells that could affect brain development are exempt from panel review, because there is little chance they would become human-like, says Carrie Wolinetz, NIH's associate director for science policy in Washington DC.

Currently, researchers use chimaeras to study early embryonic development and human diseases. But a major goal is to engineer animals to grow human organs that could then be transplanted into patients.

Unlike in the United States, it is illegal to perform such research without approval in the United Kingdom, even with private funding.

Steven Goldman, a neuroscientist at the University of Rochester in New York, says that the 2015 ban was overkill and is relieved that it will be lifted.

But Ali Brivanlou, a developmental biologist at Rockefeller University in New York City, says that the new rules should focus on limiting the percentage of the animal that becomes human instead of restricting the timing of modifications.

Bioethicist Françoise Baylis, at Dalhousie University in Halifax, Canada, worries that there are no clear guidelines on how chimaeras should be treated when used as research subjects.

These are the kinds of questions that the oversight panel will discuss when reviewing grant applications, says Wolinetz. The NIH proposal is open for public comment for 30 days, after which the agency will issue a final rule. Wolinetz hopes that it will be ready for the January 2017 grant cycle. ■

GENE EDITING

CRISPR alternative doubted

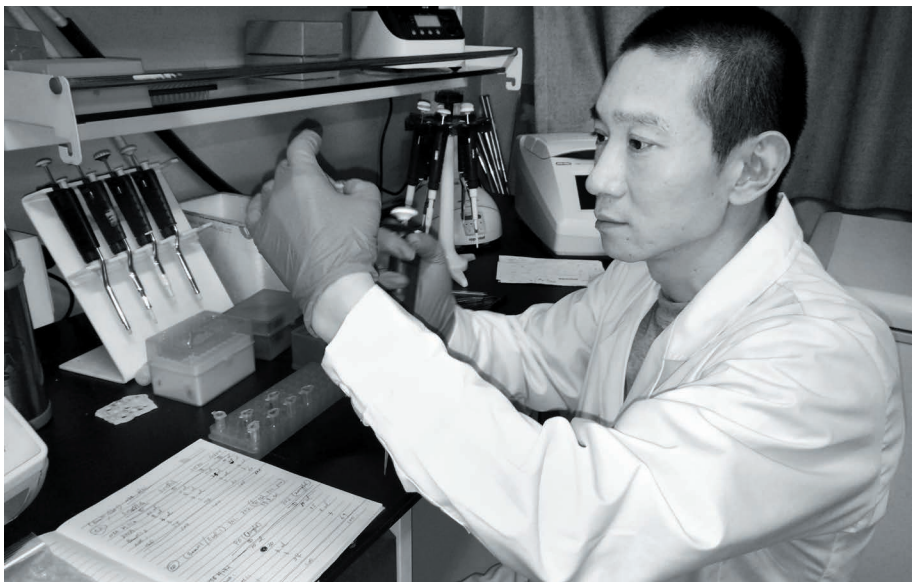
Reports of irreproducibility multiply, but author stands by his NgAgo gene-editing system.

BY DAVID CYRANOSKI, SHIJIAZHUANG, CHINA

Controversy is escalating over whether a gene-editing technique proposed as an alternative to the popular CRISPR–Cas9 system actually works.

Three months ago, Han Chunyu, a biologist at Hebei University of Science and Technology in Shijiazhuang, reported that the enzyme NgAgo can be used to edit mammalian genes. But scientists are increasingly complaining that they cannot replicate the results — although one researcher has told *Nature* that he can. *Nature Biotechnology*, which published the research, is investigating the matter.

Han says he receives dozens of harassing calls and texts each day, mocking him and telling him that his career is over — but he is convinced that the technique is sound. On 8 August, he submitted a protocol to the online genetic-information repository Addgene. He hopes that this will help efforts to reproduce his work, but other scientists say it does not clear things up.



DAVID CYRANOSKI

Han Chunyu maintains that the NgAgo enzyme can edit genes.

The stakes are high. Over the past few years, the CRISPR–Cas9 system has transformed biology. But it has also made scientists hungry to expand the gene-editing toolkit (see 'A guide to the many other ways to edit a genome'). NgAgo is one of several methods that have emerged. "A lot of us are really cheerleading and hoping that it works," says geneticist George Church of Harvard Medical School in Boston, Massachusetts.

CRISPR–Cas9 uses small genetic sequences to guide an enzyme to cut DNA in a particular location. In the *Nature Biotechnology* paper, Han's team reports using a wide variety of genetic sequences to guide NgAgo — which belongs to the Argonaute (Ago) family of proteins that others had flagged as potential gene editors — to edit eight different genes in human cells and to insert genes at specific points on chromosomes (F. Gao *et al. Nature Biotechnol.* **34,** 768–773; 2016).

NgAgo cuts only the target genes, says Han,

BEYOND CRISPR

A guide to the many other ways to edit a genome

The CRISPR–Cas9 tool enables scientists to alter genomes practically at will. It has blazed through labs around the world, finding new applications in medicine and basic research.

But the zeal with which researchers jumped on a possible new system called NgAgo earlier this year reveals an undercurrent of frustration with CRISPR–Cas9 — and a drive to find alternatives. Some are variations on the CRISPR theme; others offer new ways to edit genomes (see go.nature.com/2bbgxwb for more).

A MINI-ME

CRISPR–Cas9 may one day be used to rewrite the genes responsible for genetic diseases. But the components of the system — an enzyme called Cas9 and a strand of RNA that directs the enzyme to the desired sequence — are too large to stuff into the genome of the virus most commonly used in gene therapy to shuttle foreign genetic material into human cells.

A solution comes in the form of a mini-Cas9, which was plucked from the bacterium *Staphylococcus aureus*. It's small enough to squeeze into the virus used in one of the gene therapies currently on the market. Two groups have now used the mini-Cas9 in mice to correct the gene responsible for Duchenne muscular dystrophy.

EXPANDED REACH

Cas9 will not cut everywhere it's directed to — a certain DNA sequence must be nearby for that to happen. This demand is easily met in many genomes, but can be a painful limitation for some experiments. Researchers are looking to microbes to supply enzymes that have different sequence requirements to expand the number of sequences they can modify.

One such enzyme, called Cpf1, may become an attractive alternative. Smaller than Cas9, it has different sequence requirements and is highly specific.

Another enzyme, called C2c2, targets RNA rather than DNA — a feature that holds potential for studying RNA and combating viruses with RNA genomes.

TRUE EDITORS

Many labs use CRISPR–Cas9 only to delete sections in genes, thereby abolishing their function. "People want to declare victory like that's editing," says George Church, a geneticist at Harvard Medical School in Boston, Massachusetts. "But burning a page of the book is not editing the book."

Those who want to swap one sequence with another face a more difficult task. When Cas9 cuts DNA, the cell often makes mistakes as it

whereas CRISPR–Cas9 sometimes edits the wrong genes. And CRISPR–Cas9 requires a certain genetic sequence to be near the cutting site to initiate its activity, but NgAgo does not, which could broaden its applications, adds Han.

The initial reaction to the work in China was laudatory, including a visit to the lab by China Central Television. It was overwhelming, says Han. He doesn't like to travel and has never left China: a trip to visit a collaborator in Hangzhou in March was the first time the 42-year-old had boarded a plane. Before his paper came out, "I was completely unknown", says Han, who spoke to *Nature* at his laboratory and at a restaurant.

Doubts about the research first surfaced at the beginning of July, when Fang Shimin, a former biochemist who has become famous for exposing fraudulent scientists, wrote on his website New Threads (xys.org) that he had heard reports of failed reproduction efforts, and alleged that Han's paper was irreproducible. Criticism grew on various Chinese sites.

On 29 July, Gaetan Burgio, a geneticist at the Australian National University in Canberra, posted thorough details of his failed attempts to replicate the experiment on his blog. Normally, his posts get a few dozen hits, but this one spiked to more than 5,000.

On the same day, geneticist Lluís Montoliu at the Spanish National Centre for Biotechnology in Madrid e-mailed his colleagues at the International Society for Transgenic Technologies to recommend "abandoning any project involving the use of NgAgo". The e-mail was leaked and posted on Fang's website.

An online survey by molecular biologist Pooran Dewari of the MRC Centre of Regenerative Medicine in Edinburgh, UK, has found only 9 researchers who say that NgAgo works — and 97 who say that it doesn't. And two researchers who initially reported success with NgAgo in an online chat group now say they were mistaken.

Debojyoti Chakraborty, a molecular biologist at the CSIR-Institute of Genomics and Integrative Biology in New Delhi, says that he repeated a section of Han's paper that described using NgAgo to knock out a gene for a fluorescent protein. The glow was reduced in his cells, so Chakraborty assumed that NgAgo had disabled the gene. But DNA sequencing revealed no evidence of gene editing. Jan Winter, a PhD student in genomics at the German Cancer Research Center in Heidelberg, describes a similar experience.

JCSMR/ANU

"It's not worth pursuing. It won't surpass CRISPR, not by a long shot."
Gaetan Burgio

Han has only got the system to work on cells cultured in his laboratory. It failed in cells that he purchased, which he later found to be contaminated with *Mycoplasma* bacteria. Others might be having the same problem, he says, and some graduate students might not be being careful with reagents. Winter disagrees: "I do not think it is a problem of the scientists doing something wrong."

One researcher in China who doesn't want his name to be entangled in the controversy told *Nature* that he had tested NgAgo in a few kinds of cell and found that it was able to induce genetic mutations at the desired sites — a finding that he verified by sequencing. He adds that the process was less efficient than CRISPR–Cas9, "but, in short, it worked".

Two more Chinese scientists, who also asked not to be named, say they have initial results showing that NgAgo works but they still need to confirm with sequencing.

"It might, might work," says Burgio, "but if so, it's so challenging that it's not worth pursuing. It won't surpass CRISPR, not by a long shot."

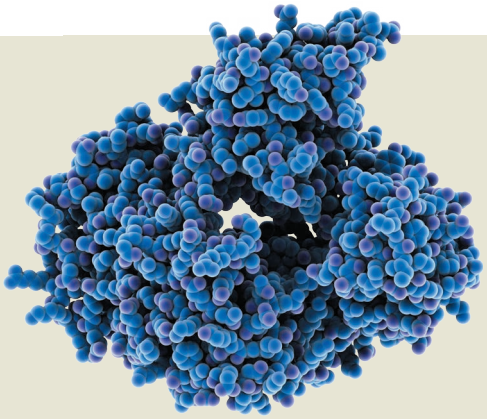
He says there is little that is new in the revised protocol on Addgene. There is a warning to maintain levels of magnesium in cells, "but that doesn't make any sense to me", he says. It also warns against *Mycoplasma* contamination. But Montoliu, who might now give NgAgo one more chance in September, doubts that this could account for all the reported problems.

The failure of NgAgo "would be disappointing", says microbiologist John van der Oost of Wageningen University in the Netherlands, a co-author of the 2014 analysis of Argonaute proteins that laid the groundwork for their use in gene editing (D. C. Swarts *et al. Nature* **507,** 258–261; 2014). "But then there is work for us left to do to see whether other Argonaute systems can get it to work somehow."

Last week, *Nature Biotechnology* sent a statement to *Nature*'s news team, which is editorially independent, saying that "several researchers" have contacted the journal to report that they cannot reproduce the results, and that "the journal is following established process to investigate the issues".

Hebei University says that it will ask Han to repeat the experiment so that it can be verified by an independent party within a month, according to Chinese state media. ■

Additional reporting by Heidi Ledford.



LAGUNA DESIGN/SPL

An argonaute protein is one of many alternatives to the CRISPR–Cas9 gene-editing system.

stitches together the broken ends. This creates the deletions that many researchers desire. But researchers who want to rewrite a DNA sequence rely on a different repair pathway that can insert a new sequence — a process that occurs at a much lower frequency than the error-prone stitching. That low efficiency poses a problem in many organisms, including some plants. "Everyone says the future is editing many genes at a time, and I think: 'We can't even do one now with reasonable efficiency'," says plant scientist Daniel Voytas at the University of Minnesota in St Paul.

But developments in the past few months have given Voytas hope. Two groups of researchers have come up with techniques that disable Cas9 then tether it to an enzyme that converts one DNA letter to another. Voytas and others are hopeful that tethering other enzymes to the disabled Cas9 will allow different sequence changes.

PURSUING ARGONAUTES

When researchers claimed in May that they could use a protein from the Argonaute family called NgAgo to slice DNA at a predetermined site without needing a guide RNA or a specific neighbouring genome sequence (F. Gao *et al. Nature Biotechnol.* **34,** 768–773; 2016), they kicked off a wave of excitement. But laboratories have so far failed to reproduce the results. Even so, there is still hope that other Argonaute proteins could provide a way forward, says genome engineer Jin-Soo Kim at the Institute for Basic Science in Seoul.

PROGRAMMING ENZYMES

Other gene-editing systems are also in the pipeline, although some have lingered there for years. For an extensive bacterial project, Church's lab did not reach for CRISPR at all. Instead, the team relied heavily on a system called lambda Red, which can be programmed to alter DNA sequences without the need for a guide RNA. But despite being studied for 13 years in Church's lab, lambda Red works only in bacteria.

Church and Feng Zhang, a bioengineer at the Broad Institute of MIT and Harvard in Cambridge, Massachusetts, say that their labs are also working on developing enzymes called integrases and recombinases for use as gene editors. "By exploring the diversity of enzymes, we can make the genome-editing toolbox even more powerful," says Zhang. "We have to continue to explore the unknown." **Heidi Ledford**

ARCHAEOLOGY

Coastal route for first Americans

Life in Canadian corridor was too late to sustain migrations of Clovis and pre-Clovis people.

BY EWEN CALLAWAY

Archaeologists need a new theory for the colonization of the Americas. Plant and animal DNA buried under two Canadian lakes squashes the idea that the first Americans travelled through an ice-free corridor that extended from Alaska to Montana.

The analysis, published online in *Nature* on 10 August and led by palaeogeneticist Eske Willerslev of the University of Copenhagen, suggests that the passageway became habitable 12,600 years ago (M. W. Pedersen *et al. Nature* http://dx.doi.org/10.1038/nature19085; 2016). That's nearly 1,000 years after the formation of the Clovis culture — once thought to be the first Americans — and even longer after other, pre-Clovis cultures settled the continents.

Some 14,000 years ago, glaciers in central Canada receded, before the appearance of Clovis people across what is now the central United States. "That coincidence seemed too powerful to ignore," says archaeologist and co-author David Meltzer of Southern Methodist University in Dallas, Texas.

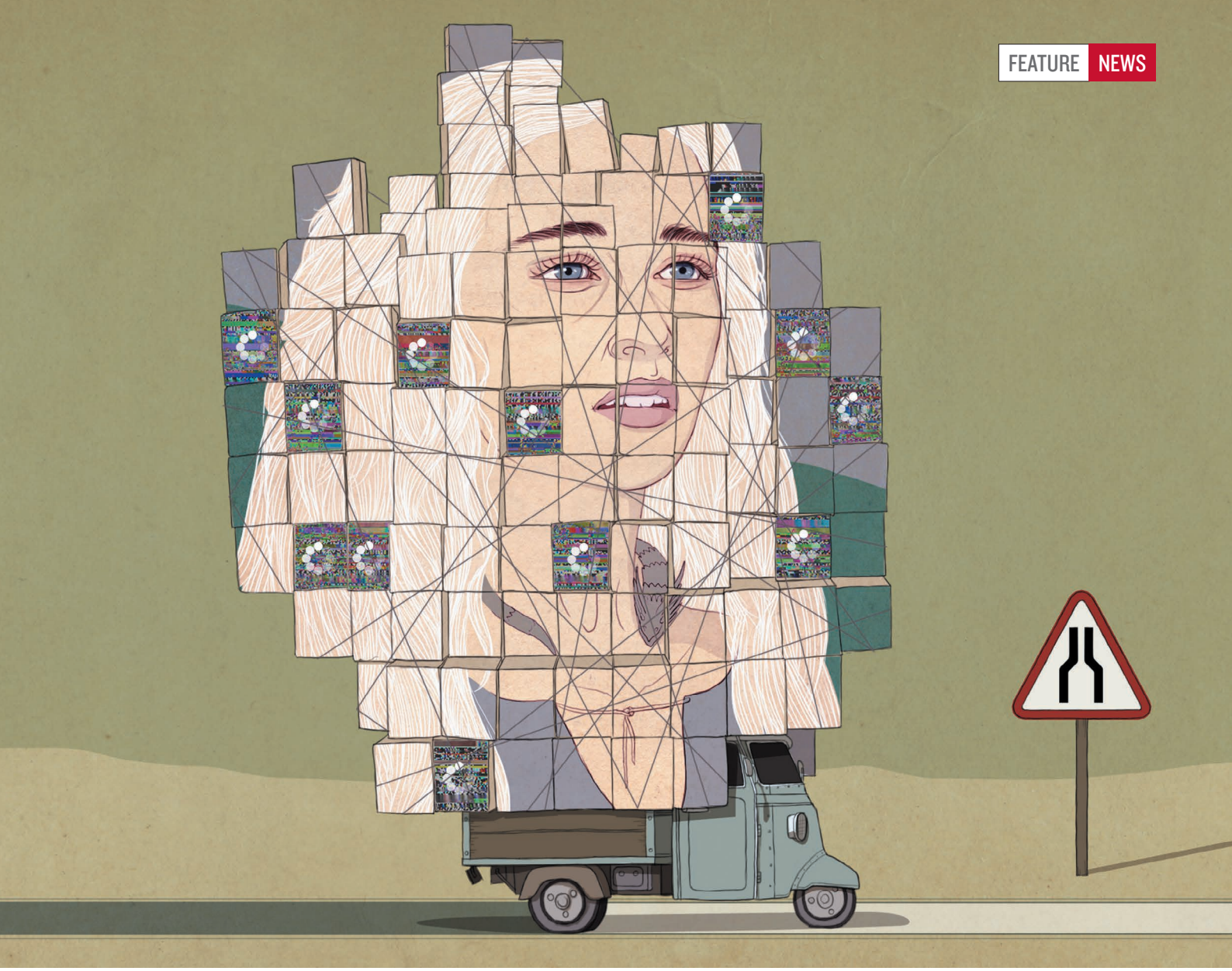
The ice-free-corridor theory began to crack in the 1990s, when researchers made a case that humans lived at Monte Verde in Chile more than 14,000 years ago. The discovery of other possible pre-Clovis sites in North America further shook the theory that Clovis people were the first Americans. But the idea that their ancestors at least trekked through the corridor persisted, says Meltzer, even though there was little consensus on when the passage opened or when it became habitable. "It's 1,500 kilometres. You can't pack a lunch and do it in a day."

"The ice-free corridor has been shown to be dead in the water."

To build a picture of the habitat as it crept out of the Ice Age, Willerslev's team analysed DNA in cores taken from beneath two lakes in what was the last stretch of the corridor to melt. The first plant life — thin grasses and sedges — dates back just 12,600 years. The region later became lusher, with sagebrush, buttercups and even roses, followed by willow and poplar trees. This habitat attracted bison first, and later mammoths, elk, voles and the occasional bald eagle. Around 11,500 years ago, the corridor began to resemble the pine and spruce boreal forests of today's landscape.

The region's bounty must eventually have tempted hunter-gatherers. But the dates rule out its use as a corridor by Clovis people and earlier groups to colonize the Americas, says Willerslev. Instead, both probably skirted the Pacific coast, perhaps by boat.

Loren Davis, an archaeologist at Oregon State University in Corvallis, agrees: "Now that the ice-free corridor has been shown to be dead in the water — no pun intended — we can start to look at something like a coastal migration route." ■



THE BANDWIDTH BOTTLENECK

Researchers are scrambling to repair and expand data pipes worldwide — and to keep the information revolution from grinding to a halt.

BY JEFF HECHT

On 19 June, several hundred thousand US fans of the television drama *Game of Thrones* went online to watch an eagerly awaited episode — and triggered a partial failure in the channel's streaming service. Some 15,000 customers were left to rage at blank screens for more than an hour.

The channel, HBO, apologized and promised to avoid a repeat. But the incident was just one particularly public example of an increasingly urgent problem: with global Internet traffic growing by an estimated 22% per year, the demand for bandwidth is fast outstripping providers' best efforts to supply it.

Although huge progress has been made since the 1990s, when early web users had to use dial-up modems and endure 'the world wide wait', the Internet is still a global patchwork built on top of a century-old telephone system. The copper lines that originally formed the system's core have been replaced by fibre-optic cables carrying trillions of bits per second between massive data centres. But service levels are much lower on local links, and at the user end it can seem like the electronic equivalent of driving on dirt roads.

The resulting digital traffic jams threaten to throttle the information-technology revolution. Consumers

ILLUSTRATION BY RICHARD WILKINSON

can already feel those constraints when mobile-phone calls become garbled at busy times, data connections slow to a crawl in crowded convention centres and video streams stall during peak viewing hours. Internet companies are painfully aware that today's network is far from ready for the much-promised future of mobile high-definition video, autonomous vehicles, remote surgery, telepresence and interactive 3D virtual-reality gaming.

That is why they are spending billions of dollars to clear the traffic jams and rebuild the Internet on the fly — an effort that is widely considered to be as crucial for the digital revolution as the expansion of computer power. Google has partnered with 5 Asian telecommunication companies to lay an 11,600-kilometre, US$300-million fibre-optic cable between Oregon, Japan and Taiwan that started service in June. Microsoft and Facebook are laying another cable across the Atlantic, to start service next year. "Those companies are making that fundamental investment to support their businesses," says Erik Kreifeldt, a submarine-cable expert at telecommunications market-research firm TeleGeography in Washington DC. These firms can't afford bottlenecks.

Laying new high-speed cable is just one improvement. Researchers and engineers are also trying several other fixes, from speeding up mobile networks to turbo-charging the servers that relay data around the world.

THE FIFTH GENERATION

For the time being, at least, one part of the expansion problem is comparatively easy to solve. Many areas in Europe and North America are already full of 'dark fibre': networks of optical fibres that were laid down by over-optimistic investors during the Internet bubble that finally burst in 2000, and never used. Today, providers can often meet rising demand simply by starting to use some of this dark fibre.

But such hard-wired connections don't help with the host of mobile phones, fitness trackers, virtual-reality headsets and other gadgets now coming online. Data traffic from mobile devices is increasing by an estimated 53% per year — most of which will end up going through mobile-phone towers, or 'base stations', whose coverage is already spotty, and whose bandwidth has to be shared by thousands of users.

The quality is spotty, as well. First-generation mobile-phone networks, introduced in the 1980s, used analogue signals and are long gone. But second-generation (2G) networks, which added digital services such as texting in the early 1990s, still account for 75% of mobile subscriptions in Africa and the Middle East, and are only now being phased out elsewhere. As of last year, the majority of mobile-phone users in Western Europe were on 3G networks, which were launched in the late 1990s to allow for more sophisticated digital services such as Internet access.

The most advanced commercial networks are now on 4G, which was introduced in the late 2000s to provide smartphones with broadband speeds of up to 100 megabits per second, and is now spreading fast. But to meet demand expected by the 2020s, say industry experts, wireless providers will have to start deploying fifth-generation (5G) technology that is at least 100 times faster, with top speeds measured in tens of billions of bits per second.

The 5G signals will also need to be shared much more widely than is currently feasible, says Rahim Tafazolli, head of the Institute for Communication Systems at the University of Surrey in Guildford, UK. "The target is how can we support a million devices per square kilometre," he says — enough to accommodate a burgeoning 'Internet of Things' that will range from networked household appliances to energy-control and medical-monitoring systems, and autonomous vehicles (see 'Bottleneck engineering').

The transition to 5G, like those to 3G and 4G before it, is being coordinated by an industry consortium that has retained the name Third Generation Partnership Project (3GPP). Tafazolli is working with this consortium to test a technique known as multiple-input, multiple-output (MIMO) — basically, a way to make each radio frequency carry many streams of data at once without letting them mix into gibberish. The idea is to put multiple antennas on both transmitter and receiver, creating many ways for signals to leave one and arrive at the other. Sophisticated signal processing can distinguish between the various paths, and extract independent data streams from each.

MIMO is already used in Wi-Fi and 4G networks. But the small size of smartphones currently limits them to no more than four antennas each, and the same number on base stations. So a key goal of 5G research is to squeeze more antennas onto both.

Big wireless companies have demonstrated MIMO with very high antenna counts in the lab and at trade shows. At the Mobile World Congress in Barcelona, Spain, in February, equipment-maker Ericsson ran live indoor demonstrations of a multiuser massive MIMO system, using a 512-element antenna to transmit 25 gigabits per second between a pair of terminals, one stationary and the other moving on rails. The system is one-quarter of the way to the 100-gigabit 5G target, and it transmits at 15 gigahertz, part of the high-frequency band planned for 5G. Japanese wireless operator NTT DoCoMo is working with Ericsson to test the equipment outdoors, and Korea Telecom is planning to demonstrate 5G services when South Korea hosts the next Winter Olympics, in 2018.

Another approach is to make the devices much more adaptive. Instead of operating on a single, hard-wired set of frequencies, a mobile device could use what is sometimes called cognitive radio: a device that uses software to switch its wireless links to whatever radio channel happens to be open at that moment. That would not only keep data automatically moving through the fastest channels, says Tafazolli, but also improve network resilience by finding ways to route around failure points. And, he says, it's much easier to upgrade performance by replacing software than by replacing hardware.

Meanwhile, a crucial policy challenge for the 5G transition is finding a radio spectrum that offers adequate bandwidth and coverage. International agreements have already allocated almost every accessible frequency to a specific use, such as television broadcasting, maritime navigation or even radio astronomy. So final changes will have to wait for the 2019 World Radiocommunication Conference. But the US Federal Communications Commission (FCC) is trying to get a head start by auctioning off frequencies below 1 gigahertz to telecommunications companies. Once reserved for broadcast television because they are better than higher frequencies at penetrating walls and other obstructions — but no longer needed after television's shift to digital — these low frequencies are particularly attractive for serving sparsely populated areas, says Tafazolli: only a few base stations would be required to provide broadband service to households and driving data to autonomous cars on motorways.

DIGITAL TRAFFIC JAMS THREATEN TO THROTTLE THE INFORMATION-TECHNOLOGY REVOLUTION.

Other bands in the 1–6-gigahertz range could be opened up for 5G use as 2G and 3G technologies are phased out. But the best hope for dense urban areas is to exploit frequencies above 6 gigahertz, which are currently little-used because they have a very short range. That would require 5G base stations up to every 200 metres in dense urban areas, one-fifth the spacing typical of urban 4G networks. But the FCC considers the idea promising enough that on 14 July, it formally approved opening these frequencies for high-speed, fast-response services. Ofcom, the UK regulatory body, is considering similar steps.

Companies are particularly interested in these higher frequencies as a way to extend 5G technology for other uses. In the United States, wireless carrier Verizon and a consortium

NATURE.COM
To listen to a podcast about the bandwidth challenge, visit:
go.nature.com/2axbk00

of equipment-makers including Ericsson, Cisco, Intel, Nokia and Samsung have tested 28-gigahertz transmission at sites in New Jersey, Massachusetts and Texas. The system uses 5G technology to deliver data at 1 gigabit per second, and Verizon is adapting it for use in fixed wireless connections to homes, which it plans to test next year. The company has been pushing fixed wireless as an alternative to wired connections, because connection costs are much lower.

BIGGER PIPES

"When I take out my cell phone, everyone thinks of it as a wireless communications device," says Neal Bergano, chief technology officer of TE SubCom, a submarine-cable manufacturer based in Eatontown, New Jersey. Yet that is only part of the story, he says: "Users are mobile, but the network isn't mobile." When someone uses their phone, its radio signal is converted at the nearest base station to an optical signal that then has to travel to its destination through fixed fibre optics.

These flexible glass data channels have been the backbone of the global telecommunications network for more than a quarter of a century. Nothing can match their bandwidth: today, a single hair-thin fibre can transmit 10 terabits (trillion bits) per second across the Atlantic. That is the equivalent of 25 double-layer Blu-ray Discs per second, and is 30,000 times the capacity of the first transatlantic fibre cable, laid in 1988. Much of that increase came when engineers learned how to send 100 separate signals through a single fibre, each at its own wavelength. But as traffic continues to increase over heavily used routes, such as New York to London, that approach is coming up against some hard limits: distortion and noise that inevitably build up as light passes along thousands of kilometres of glass have made it effectively impossible to send more than 100 gigabits per second on a single wavelength.

To overcome that limit, manufacturers have developed a new type of fibre. Whereas standard fibres send the light through a 9-micrometre-wide core of ultrapure glass running down the middle, the newer design spreads the light over a larger core area at lower intensity, reducing noise. The trade-off is that the new fibres are more sensitive to bending and stretching, which can introduce errors. But they work very well in submarine cables, because the deep sea provides a benign, stable environment that puts little strain on the fibre.

Last year, networking-systems firm Infinera in Sunnyvale, California, sent single-wavelength signals at 150 gigabits per second through a large-area fibre for 7,400 kilometres — more than 3 times the distance possible with a standard fibre, and easily enough to cross the Atlantic. They also transmitted 200-gigabit-per-second signals a shorter distance.

The highest-capacity commercial submarine cable now in service is the 60-terabit-per-second FASTER system that opened in June between Oregon and Japan. It sends 100-gigabit-per-second signals on 100 wavelengths in each of 6 pairs of large-core fibres. But in late May, Microsoft and Facebook jointly announced plans to beat it with MAREA: a large-area fibre cable spanning the 6,600 kilometres between Virginia and Spain. When completed in October 2017, the cable will link the two companies' data centres on opposite sides of the Atlantic at 160 terabits per second.

Another approach to reducing performance-limiting noise was demonstrated last year by a group at the University of California, San Diego. Fibre-optic systems normally use separate lasers for each wavelength, but tiny, random variations can generate noise. Instead, the group used a technique known as a frequency comb to generate a series of uniformly spaced wavelengths from a single laser (E. Temprana *et al*. *Science* **348,** 1445–1448; 2015). "It worked like a charm" to reduce noise, says group member Nikola Alic, an electrical engineer. With further development, he says, the approach could double the data rate of fibre-optic systems.

TIME OF FLIGHT

Impressive bandwidth is useful, but promptness also matters. Human speech is so sensitive to interruption that a delay of one-quarter of a second can disturb a phone or video conversation. Video requires a fixed frame rate, so streaming video stalls when its input queue runs dry. To overcome such problems, FCC rules allow special codes that give priority passage for packets of data carrying voice calls or video frames, so that they flow quickly and uniformly through the Internet.

New and emerging services including telerobotics, remote surgery, cloud computing and interactive gaming are also sensitive to network responsiveness. The time it takes for a signal to make a round trip between two terminals, often called latency, depends largely on distance — a reality that shapes the geography of the Internet. Even though data travel through fibre-optic cable at 200,000 kilometres per second, two-thirds the velocity of light in the open air, a person tapping a key in London would still need 86 milliseconds to get a response from a data centre in San Francisco, 8,600 kilometres away — a delay that would make cloud computing crawl.

Emerging mobile applications require both broad bandwidth and low latency. Autonomous cars, for example, need real-time data on their environment to warn them about hazards, from potholes to accidents ahead. Conventional cars are becoming wireless nerve centres, needing low latency for 'hands-free' voice-control systems.

A potentially huge challenge is the emergence of 3D virtual-reality systems. Interactive 3D gaming requires data to travel at 1 gigabit per

BOTTLENECK ENGINEERING

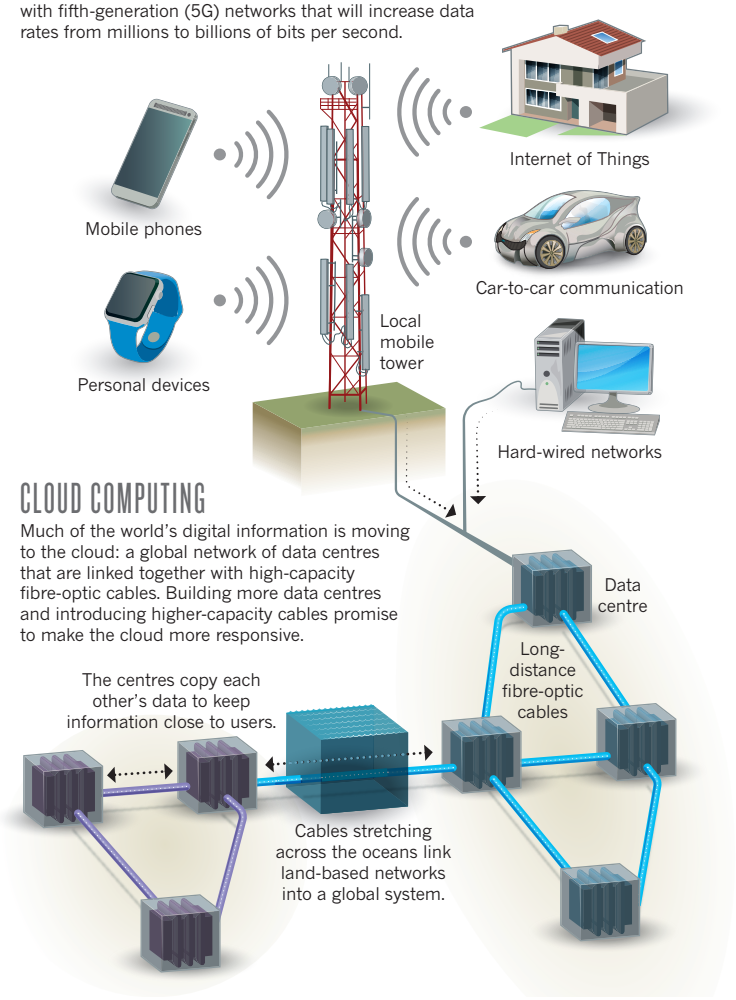
The Internet was built on a century-old telephone system, leaving many choke points that have to be eliminated to keep the bits flowing.

MOBILE EVERYTHING

Demand for wireless connections is exploding, with ever more devices coming online. Engineers hope to meet that demand with fifth-generation (5G) networks that will increase data rates from millions to billions of bits per second.

CLOUD COMPUTING

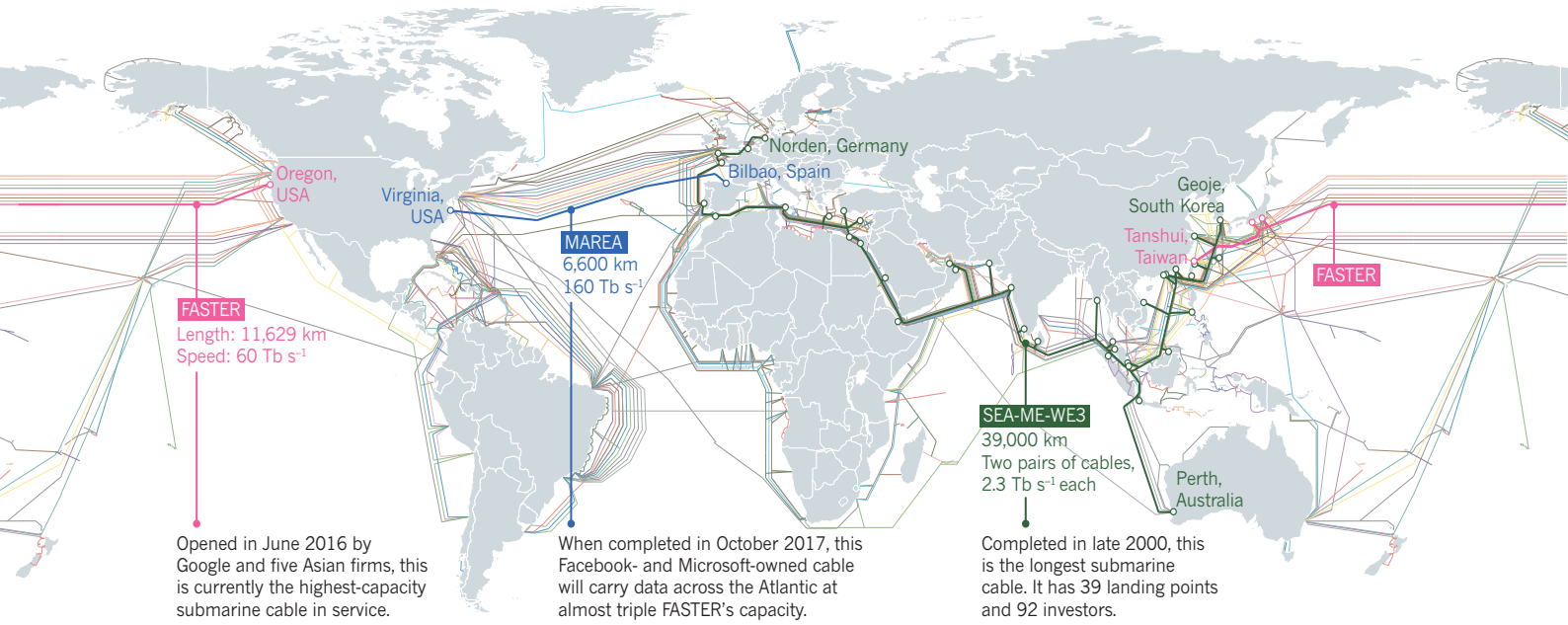
Much of the world's digital information is moving to the cloud: a global network of data centres that are linked together with high-capacity fibre-optic cables. Building more data centres and introducing higher-capacity cables promise to make the cloud more responsive.



NIK SPENCER/*NATURE*

THE SUBMARINE WEB

Much of the world's Internet traffic passes under the oceans, through fibre-optic cables that can run along the sea bed for thousands of kilometres. Companies are constantly laying more and better cables.



SOURCE: TELEGEOGRAPHY

second — 20 times the speed of a typical video feed from a Blu-Ray Disc. But most crucially, the image must be rewritten at least 90 times per second to keep up with users turning their heads to watch the action, says computer scientist David Whittinghill of Purdue University in West Lafayette, Indiana. If the data stream slips behind, the user gets motion sickness. To keep that from happening, Whittinghill has installed a special 10-gigabit-per-second fibre line to his virtual-reality lab.

To speed up responses, big Internet companies such as Google, Microsoft, Facebook and Amazon store replicas of their data in multiple server farms around the world, and route queries to the closest. Video cached at a local data centre is what allows viewers to fast-forward as if the file was stored on a home device, says Geoff Bennett, director of solutions and technology for Infinera. But the proliferation of these data centres is also one of the biggest drivers of bandwidth demand, he says: vendors' efforts to synchronize private data centres around the world now consume more bandwidth than public Internet traffic. The Microsoft–Facebook cable is being built expressly for this purpose (see 'The submarine web').

So far, most data centres are where the customers and cables are: in North America, Europe and east Asia. "Many parts of the world still rely on remote access to content that is not stored locally," says Kreifeldt. South America has few data centres, he says, so much of the content comes from well-wired Miami, Florida: traffic between Chile and Brazil might be routed through Miami to save money, but at a cost in latency. The same problem plagues the Middle East, where 85% of international traffic must travel to centres in Europe. That is changing, says Kreifeldt, but progress is slow. Amazon Web Services launched its first cloud data centre in India this year, in Mumbai; it has had a similar centre in São Paulo, Brazil, since 2011.

> "USERS ARE MOBILE, BUT THE NETWORK ISN'T MOBILE."

INTERNAL COMMUNICATIONS

Bandwidth is also crucial on the very smallest scale: on and between the chips in the banks of servers in a data centre. Expanding the flow here can help information to move more quickly within the data centres and get out to users faster. Chip clock speeds — how fast the chip runs — flat-lined at a few gigahertz several years ago, because of heating problems. The most practical way to speed up processors significantly is to divide the operations that they perform between multiple 'cores': separate microprocessors operating in parallel on the same chip. That requires high-speed connections within the chip — and one way to make them is with light, which can move data faster than electrons can.

The biggest obstacle has been integrating microscale optics with silicon electronics. After years of research on 'silicon photonics', engineers have yet to find a way to efficiently generate light from silicon, a key step in optical information processing. The best semiconductor light sources, such as indium phosphide, can be bonded to silicon chips, but are very difficult to grow directly on silicon, because their atoms are spaced differently. Optical and electronic components have been integrated on indium phosphide, but so far only on a small scale.

In an effort to scale up photonic integration to a commercial level, the United States last year launched the American Institute for Manufacturing Integrated Photonics in Rochester, New York, which is supported by $110 million from federal agencies and $502 million from industry and other sources. Its target is to develop an efficient technology to make integrated photonics for high-speed applications, including optical communications and computing.

Separately, a Canadian-funded team earlier this year demonstrated a photonic integrated circuit with 21 active components that could be programmed to perform 3 different logic functions (W. Liu *et al. Nature Photon.* **10,** 190–195; 2016). That's an important step for photonic microprocessors, comparable in complexity to the first programmable electronic chips that opened the door to microcomputers. "Compared to current electronics, it's simple, but compared to photonic integrated circuits it is quite complicated," says study co-author Jianping Yao, an electrical engineer at the University of Ottawa in Canada.

Further development could lead to varied applications. For example, Yao says that after the chip is optimized for manufacture, it could convert a 5G smartphone signal received at a base station into an analogue optical signal, which could be transmitted by fibre optics to a central facility, and then digitized.

The quest for faster chips, like other parts of the Internet problem, is a daunting challenge. But researchers such as Bergano see a lot of potential for improvements. After 35 years of working on fibre optics, he says, "I remain a complete optimist when I think about the future." ■

Jeff Hecht *is a freelance writer in Auburndale, Massachusetts.*

COMMENT

MOHD SAMSUL MOHD SAID/GETTY

A container of seized African elephant tusks in Malaysia.

The ravages of guns, nets and bulldozers

The threats of old are still the dominant drivers of current species loss, indicates an analysis of IUCN Red List data by **Sean Maxwell** and colleagues.

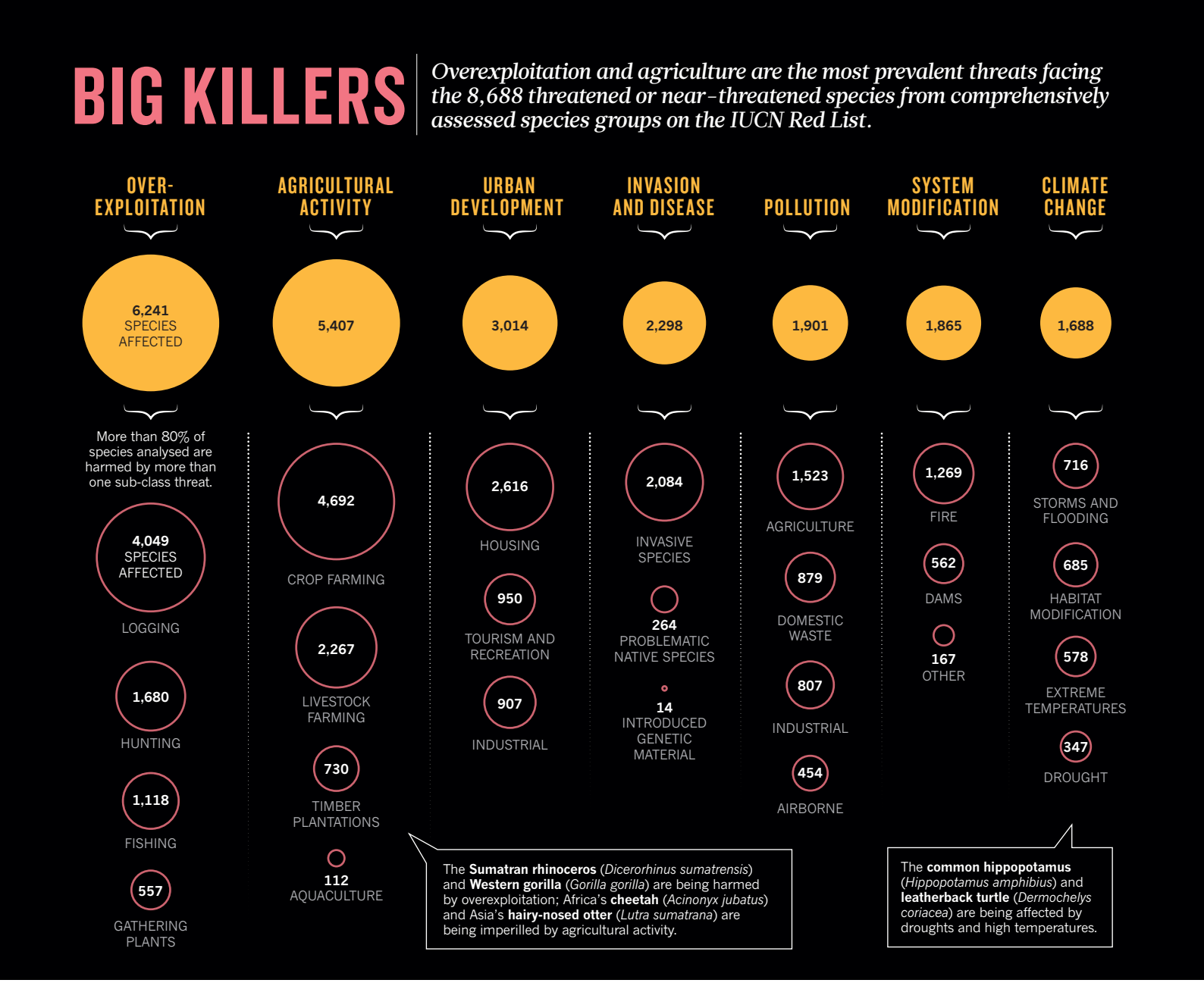
There is a growing tendency for media reports about threats to biodiversity to focus on climate change.

Here we report an analysis of threat information gathered for more than 8,000 species. These data revealed a contrasting picture. We found that by far the biggest drivers of biodiversity decline are overexploitation (the harvesting of species from the wild at rates that cannot be compensated for by reproduction or regrowth) and agriculture (the production of food, fodder, fibre and fuel crops; livestock farming; aquaculture; and the cultivation of trees).

Early next month, representatives from government, industry and non-governmental organizations will define future directions for conservation at the World Conservation Congress of the International Union for Conservation of Nature (IUCN). High on the agenda for political leaders, non-governmental organizations, conservationists and many others will be taking steps to turn the 2015 Paris climate agreement into action. It is also crucial that the World Conservation Congress delegates — and society in general — ensure that efforts to address climate change do not overshadow more immediate priorities for the survival of the world's flora and fauna.

ON THE LIST

Since 2001, the categories and criteria of the IUCN Red List of Threatened Species — a standard for the evaluation of extinction ▶



risk — have guided assessments, now for 82,845 species. Assessors assign species to categories, including 'near-threatened', 'vulnerable', 'endangered' or 'critically endangered' depending on their population size; past, current and projected population trends; geographic range and other symptoms of extinction risk. Species in the latter three groups are collectively referred to as 'threatened'.

To assess the relative prevalence of current hazards to biodiversity, we quantified threat information for 8,688 near-threatened or threatened species belonging to species groups in which all known species have been assessed (for complete list of taxa included, see Supplementary Information; go.nature.com/2ajen88).

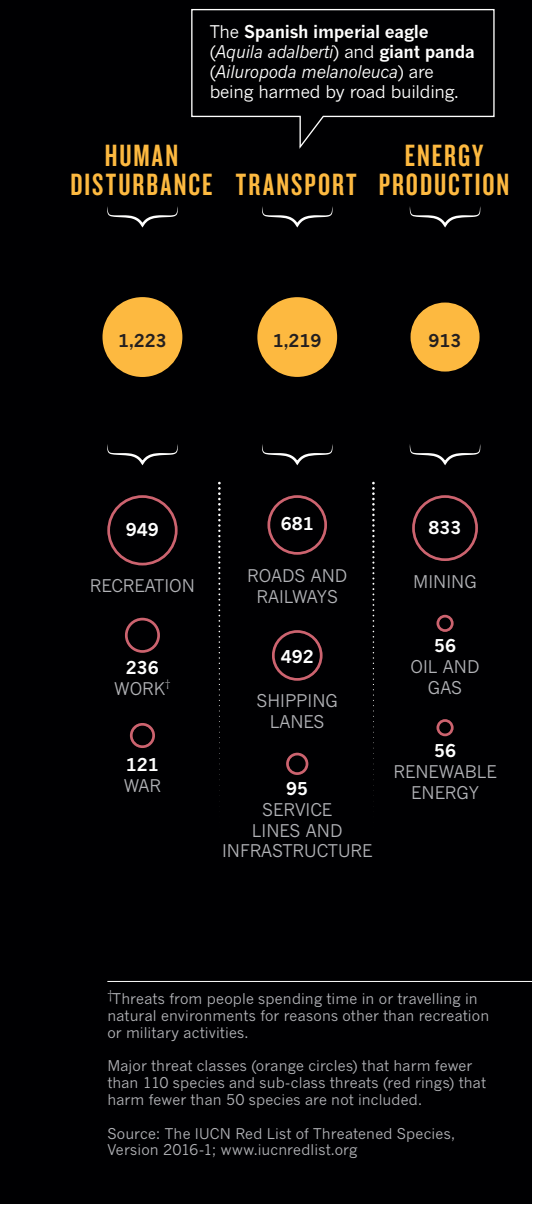
The basic message emerging from these data is that whatever the threat category or species group, overexploitation and agriculture have the greatest current impact on biodiversity (see 'Big killers').

Of the species listed as threatened or near-threatened, 72% (6,241) are being overexploited for commerce, recreation or subsistence.

The Sumatran rhinoceros (*Dicerorhinus sumatrensis*), Western gorilla (*Gorilla gorilla*) and Chinese pangolin (*Manis pentadactyla*, a scaly mammal), for instance, are all illegally hunted as a result of high market demand for their body parts and meat. These are just three of the more than 2,700 species affected by hunting or fishing, or by people collecting live specimens for the pet trade. At the same time, unsustainable logging is contributing to the decline of more than 4,000 forest-dependent species, such as the Bornean wren-babbler (*Ptilocichla leucogrammica*), India's Nicobar shrew (*Crocidura nicobarica*), and the Myanmar snub-nosed monkey (*Rhinopithecus strykeri*).

The expansion and intensification of agricultural activity is imperilling 5,407 species — 62% of those listed as threatened or near-threatened. Africa's cheetah (*Acinonyx jubatus*), Asia's hairy-nosed otter (*Lutra sumatrana*) and South America's huemul deer (*Hippocamelus bisulcus*) are among more than 2,300 species affected by livestock farming and aquaculture. And the Fresno kangaroo rat (*Dipodomys nitratoides*) and the African wild dog (*Lycaon pictus*) are two of more than 4,600 species currently under threat from land modification associated with the production of food, fodder or fuel crops.

Meanwhile, anthropogenic climate change — including increases in storms, flooding, extreme temperatures or drought that exceed background variability, as well as sea-level rise — is currently affecting 19% of species listed as threatened or near-threatened. Hooded seals (*Cystophora cristata*) are among the 1,688 species affected. These have dropped in abundance by 90% in the northeastern Atlantic Arctic over the past few decades



†Threats from people spending time in or travelling in natural environments for reasons other than recreation or military activities.

Major threat classes (orange circles) that harm fewer than 110 species and sub-class threats (red rings) that harm fewer than 50 species are not included.

Source: The IUCN Red List of Threatened Species, Version 2016-1; www.iucnredlist.org

GRAPHIC BY WES FERNANDES/*NATURE*

as a result of extensive declines in regional sea ice, and so in the availability of sites for resting and raising pups.

DATA LIMITATIONS

There are three obvious difficulties in interpreting the Red List data.

First, the patterns we report here do not necessarily extend to taxonomic groups that haven't been monitored. The comprehensively assessed groups included here are not a random sample from the tree of life[1], but those that are generally better-studied. All known bird species have been assessed, for instance. But information on extinction risk has been gathered for only some 0.1% of the more than 50,000 species of fungi thought to exist.

A second potential limitation of our analysis is that it treats threats as discrete when, in fact, hazards rarely affect organisms in isolation. Agriculture is a major driver of greenhouse-gas emissions, for example. And new roads to enable agricultural expansion can increase bush-meat harvesting, the incidence of forest fires and habitat fragmentation[2]. In fact, more than 80% of the species included in our analysis are affected by more than one major threat.

Finally, the balance of threats driving extinction risk for many of the world's species will change, even over the next few decades[3]. For Red List assessments, the impacts of future threats (including climate change) in reducing a species' population size are projected across three generations or over a ten-year period — whichever is longer. Hence, unless the species being assessed is long-lived (with an expected lifespan of 30–50 years, say), projections cover a period during which the effects of climate change, in particular, will be relatively modest.

"Climate change will become an increasingly dominant problem in the biodiversity crisis."

Yet we do not think that any of these caveats alter the overall message. Because agricultural activity and overexploitation tend to occur in fertile places with naturally high levels of biodiversity[4], the patterns emerging from our analysis probably extend to many of the other species that have not yet been assessed. Also, until a better understanding is obtained of how threats act additively, synergistically or antagonistically, a pragmatic course of action is to limit those impacts that are currently harming the most species[5]. Finally, studies have shown Red List categorizations reflecting projected extinction risk from climate change to be more robust than was previously thought[6].

WHAT NEXT?

Of all the plant, amphibian, reptile, bird and mammal species that have gone extinct since AD 1500, 75% were harmed by overexploitation or agricultural activity or both (often in combination with the introduction of invasive alien species[7]). Climate change will become an increasingly dominant problem in the biodiversity crisis[3]. But human development and population growth mean that the impacts of overexploitation and agricultural expansion will also increase.

The aim of the World Conservation Congress is to translate sustainable development and carbon neutrality agreements into action. We urge delegates to focus on proposing and funding actions that prioritize the biggest current threats to biodiversity.

Thankfully, there are effective tools and approaches to alleviate harm caused by overexploitation and agricultural activities[8]. These include the development and governance of sustainable harvest regimes; the enforcement of hunting regulations and no-take marine protected areas; the maintenance of international policy mechanisms; such as the Convention on International Trade in Endangered Species; and public education (for instance, on where ivory comes from) to reduce demand. Also powerful are the establishment of protected areas to safeguard key biodiversity areas[9]; the management of agricultural systems in ways that allow threatened species to persist within them; the regulation of pesticide and fertilizer use; the certification of agricultural sustainability; and the reduction of food waste, for example, using urban food-transfer programmes.

Crucially, ensuring that overexploitation and agricultural activities today do not compromise ecosystems tomorrow will help to ameliorate the challenges presented by impending climate change. Healthy ecosystems are better repositories for carbon. They are also more likely to provide the physical connectivity and genetic diversity needed to enable species to adapt to the large shifts in climate expected later this century[10].

Conservationists, weary of tackling herculean, long-standing problems, could be forgiven for being drawn to newer ones. Nonetheless, we appeal to all concerned with the sustainability of life on Earth to take stock of the current balance of threats — and refocus their efforts on the enemies of old. ■

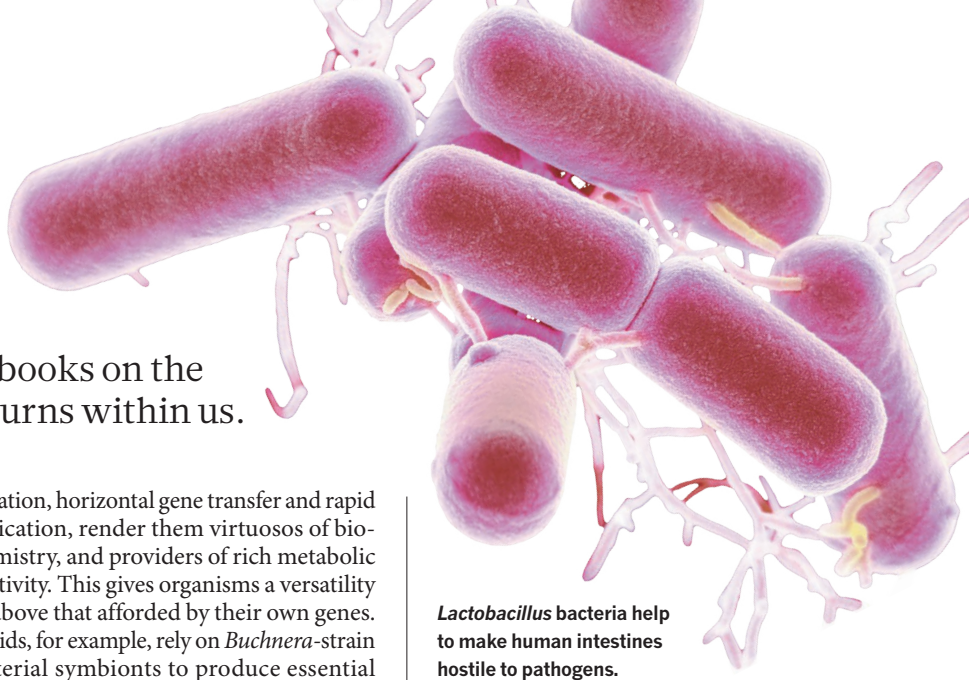
Sean L. Maxwell *is a PhD student in the School of Geography, Planning and Environmental Management, University of Queensland, Brisbane, Australia.* **Richard A. Fuller** *is an associate professor in the School of Biological Sciences, University of Queensland, Brisbane, Australia.* **Thomas M. Brooks** *is head of science and knowledge at the International Union for Conservation of Nature, Gland, Switzerland.* **James E. M. Watson** *is associate professor in the School of Geography, Planning and Environmental Management, University of Queensland, Brisbane, Australia, and director of the Science and Research Initiative at the Global Conservation Program, Wildlife Conservation Society, New York, USA.*
e-mail: smaxwell@uq.edu.au

1. Brooks, T. M. *et al. Sci. Data* **3,** 160007 (2016).
2. Laurance, W. F., Goosem, M. & Laurance, S. G. W. *Trends Ecol. Evol.* **24,** 659–669 (2009).
3. Foden, W. B. *et al. PLoS ONE* **8,** e65427 (2013).
4. Anderson, B. J. *et al. J. Appl. Ecol.* **46,** 888–896 (2009).
5. Côté, I. M., Darling, E. S. & Brown C. J. *Proc. R. Soc. B* **283,** 20152592 (2016).
6. Keith, D. A. *et al. Conserv. Biol.* **28,** 810–819 (2014).
7. Bellard, C., Cassey, P. & Blackburn, T. M. *Biol. Lett.* **12,** 20150623 (2016).
8. Hayward, M. W. *Biodivers. Conserv.* **20,** 2563–2573 (2011).
9. Watson, J. E. M., Dudley, N., Segan, D. B. & Hockings, M. *Nature* **515,** 67–73 (2014).
10. Martin, T. G. & Watson, J. E. M. *Nature Clim. Change* **6,** 122–124 (2016).

MICROBIOLOGY

Mob rule

Adrian Woolfson examines four books on the microbiological universe that churns within us.



***Lactobacillus* bacteria help to make human intestines hostile to pathogens.**

SPL

In the early 1990s, molecular biologist Sydney Brenner gave a talk in Cambridge, UK, in which he espoused the merits of sequencing the human genome to fully characterize the human "gene kit". Several years later, in 2001, the first draft sequence of the human genome was released. The assumption was that human form, function and dysfunction would be reduced to a finite and tractable problem. Over time, this vision has been eroded by the discovery of successive Russian-doll-like levels of informational and regulatory complexity, from epigenetics to microRNAs. Genomic protein-encoding genes may represent the surface of a much deeper problem.

The latest assault on Brenner's model of organismal form and function has come from an unexpected quarter. It seems that, instead of being self-contained, the contents of the human gene kit are generously supplemented by a plethora of extraneous components. These riches come from the topsy-turvy world of microorganisms, symbionts whose products bolt onto the more modest collection furnished by their hosts. The implications of this extra informational dimension, and how it interweaves with our genes, are explored in four new books.

"Our minuscule passengers act like puppet masters, manipulating how we think, feel and act."

In his compelling *I Contain Multitudes*, science writer Ed Yong plunges into the Alice in Wonderland shadow world of the microbes that live in and on us. As he reminds us, the 30 trillion cells in the human body are effortlessly outnumbered by the 39 trillion or so microbial cells that lurk within it. Our own genomes muster 20,000 protein-encoding genes; our uninvited guests may collectively field an impressive 10 million. We know this thanks to metagenomics — the method of sequencing short, species-specific stretches of RNA, pioneered by biophysicist Carl Woese in the late 1960s — which helps to define the genomic architecture of our microbial communities.

Bacteria confer unique properties on their hosts. Their collective genes, and capacity for rapid evolution through high rates of mutation, horizontal gene transfer and rapid replication, render them virtuosos of biochemistry, and providers of rich metabolic creativity. This gives organisms a versatility far above that afforded by their own genes. Aphids, for example, rely on *Buchnera*-strain bacterial symbionts to produce essential amino acids absent from the phloem sap that is the insects' food. Such relationships led US biologist Ivan Wallin in 1927 to describe symbiosis as an engine of novelty that enables bacteria to transform their host species.

Whereas scientists from germ-theory pioneer Louis Pasteur to penicillin-developer Howard Florey have taught us to fear microbes, Yong argues that we must nurture them, appreciating that they may help us to develop into what we are. The human microbiome should be viewed as a distributed organ, performing functions as essential as those of our liver, lungs or kidneys.

Intriguingly, Yong argues that human immune cells are akin more to park rangers than to xenophobes, carefully wrangling the microbial zoo, modulating its population dynamics and responding to its chatter. The degradation and collapse of coral reefs in warm, acidic waters is due not only to direct effects of global warming, but also to the disruption of relationships in microbial communities. Likewise, Yong suggests that some human diseases result from alterations to bacterial community dynamics, triggering abnormalities in internal microbial ecology and cooperativity. An example of this is obesity, which seems, in part, to result from an imbalance of gut microbes. Obese individuals have more bacteria from the phylum Firmicutes and fewer from the genus *Bacteroides* than lean ones, and a relative lack of *Akkermansia muciniphila*. It was shown in 2013 that microbes from lean mice can make obese mice lose weight (A. Everard *et al. Proc. Natl Acad. Sci. USA* **110,** 9066–9071; 2013).

Yong goes on to explain how the dialogue between cells and resident microbes may affect organismal development. Hawaiian bobtail squid (*Euprymna scolopes*) adopt their mature form only in the presence of the luminescent bacterium *Vibrio fischeri*, which colonizes the squid's light organ. Human breast milk contains indigestible oligosaccharides, the favoured food of *Bifidobacterium longum infantis*, which releases short-chain fatty acids that influence the permeability of an infant's gut cells.

In *The Human Superorganism*, immunotoxicologist Rodney Dietert goes further, asserting that *Homo sapiens* is a superorganism containing thousands of microbial species. He argues that the biology of microbes will eventually challenge our view of what it means to be human, and lead to the identification of therapeutic agents. In his vision, humans are "microbial storage machines" designed to pass microorganisms to future generations. Our "second genome" — the genes encoded by our microbiome — resides in a thriving bacterial community that he compares to the diversity of a tropical rainforest. Even in the age of genome-editing tools such as CRISPR, it remains challenging to modify the human genome. Dietert is astute, however, in suggesting that microbial genomes could be engineered to introduce functionalities and tackle human diseases. The ability of microbial metabolites to manipulate the expression of human genes has already been established: sodium butyrate, for example, helps to control the

I Contain Multitudes: The Microbes Within Us and a Grander View of Life
ED YONG
Ecco: 2016.

The Human Superorganism: How the Microbiome Is Revolutionizing the Pursuit of a Healthy Life
RODNEY DIETERT
Dutton: 2016.

This Is Your Brain on Parasites: How Tiny Creatures Manipulate Our Behavior and Shape Society
KATHLEEN MCAULIFFE
Houghton Mifflin Harcourt: 2016.

The Mind–Gut Connection: How the Hidden Conversation Within Our Bodies Impacts Our Mood, Our Choices, and Our Overall Health
EMERAN MAYER
Harper Wave: 2016.

switch from embryonic to fetal haemoglobin.

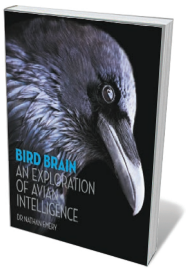
Not content to cruise around their luxury human condos, microorganisms also hack into our nervous systems. In her eye-opening, entertaining and slightly disconcerting *This Is Your Brain on Parasites*, journalist Kathleen McAuliffe contends that our minuscule passengers act like puppet masters, manipulating how we think, feel and act. I will think of cats differently now that I am aware that their parasite *Toxoplasma gondii* may have the ability to affect human behaviour, and is implicated in mental illnesses such as schizophrenia. Men harbouring this parasite are, furthermore, more inclined to break rules, and are more reserved and suspicious. McAuliffe ingeniously suggests that the psychoactive chemicals produced by microbes could be used to develop mind-altering medicines.

Focusing on how the microbiome may cause chronic conditions such as persistent pain and irritable bowel syndrome, gastroenterologist Emeran Mayer's *The Mind–Gut Connection* depicts the brain, the gut and its microorganisms as a unitary structure tightly knit, anatomically and chemically. He asserts, albeit with rudimentary evidence, that the enteric nervous system — the mesh of neurons that governs the gastrointestinal system — functions as a mini-brain, relaying sensory information from the gut to the central nervous system. It was fascinating to learn that microbes contain ancient versions of many signalling peptides and hormones found in the human alimentary tract, including noradrenaline, serotonin and endorphins. That may argue in favour of his thesis. Mayer speculates that early programming errors in the putative brain–gut–microbiome axis can result in medical conditions that might benefit from treatment with probiotics.

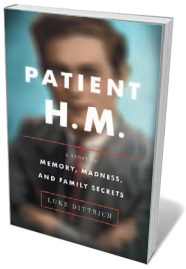
We are descended from microbes, have evolved around them, and incorporate elements of them into our cells. Microbiome profiling is certain to become as routine as blood testing, and the extensive treasure chest of bacterial molecules will doubtless be used to change the way we are. Our microbial companions may even influence our responses to important medicinal agents, such as the anti-PD-L1 and anti-CTLA4 drugs that reinvigorate the immune systems of people with cancer. Several regional initiatives, including the US Human Microbiome Project and National Microbiome Initiative, have been established to study the human microbiome. The complexities of cataloguing, mapping and characterizing microbial biology on a worldwide scale promise to make sequencing the human genome look easy. A global programme seems to beckon. ■

Adrian Woolfson *is the author of* Life Without Genes.
e-mail: adrianwoolfson@yahoo.com

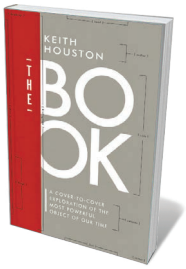
Books in brief



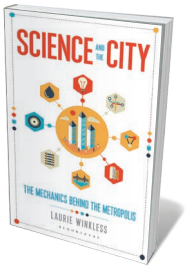
Bird Brain: An Exploration of Avian Intelligence
Nathan Emery IVY *(2016)*
Cognitive biologist Nathan Emery has been on the cutting edge of research into avian intelligence since the 1990s. In this sparkling, superbly illustrated summation of the cognitive science, ethology and hot debates, Emery encapsulates the "feathered ape". He compares the avian brain to the mammalian to reveal functional similarities in disparate anatomies (likened to fruitcake and layer cake, respectively) and tours spatial memory, migratory sense, tool use and more. From the wattle-bopping of black grouse (*Tetrao tetrix*) to the dung baiting of burrowing owls (*Athene cunicularia*), a masterful explication.



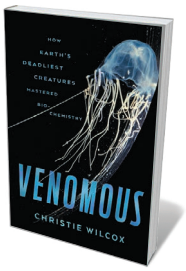
Patient H.M.: A Story of Memory, Madness, and Family Secrets
Luke Dittrich RANDOM HOUSE *(2016)*
In 1953, experimental surgery left Henry Molaison with severe amnesia; he became 'HM', a star patient studied by neuroscientist Suzanne Corkin for almost 50 years (see D. Draaisma *Nature* **497,** 313–314; 2013). Luke Dittrich offers a very different perspective — he is the grandson of William Scoville, the lobotomist who operated on Molaison. Dittrich fleshes out the official account with nuanced biographies of the troubled Scoville and profoundly damaged Molaison, revelatory conversations with Corkin and accounts of behind-the-scenes scientific scuffles. Disturbing and illuminating.



The Book
Keith Houston W. W. NORTON *(2016)*
The physical book has reigned as an agent of culture for 1,500 years. Keith Houston's deft history of the object wraps entire civilizations into the telling, propelling us through the evolution of writing, printing, binding and illustration with gusto. The material innovations dazzle, from papyrus, vellum and paper (dating to second-century AD China) to the spattered path of inks. Equally gripping is the trajectory of production technologies, as the finical skill of scribes gives way to Johannes Gutenberg's printing revolution and, ultimately, the streamlined wonders of modern lithography.



Science and the City: The Mechanics Behind the Metropolis
Laurie Winkless BLOOMSBURY SIGMA *(2016)*
'Up', 'Switch', 'Wet': physicist Laurie Winkless's chapter headings hint at a briskly bouncy ride ahead in this primer on the science embedded in cities. And so it proves, as she ponders wind-confusing skyscraper design, water-supply technologies such as "fog-sucking nets" and 3D-printed bridges. Perhaps most engrossing is her evocation of how modern subway systems are built — by delicately 'threading the needle' through dense subterranean convolutions. The thickets of subheadings and bolded-up key terms may irk, but the witty Winkless has done her homework.



Venomous
Christie Wilcox FARRAR, STRAUS AND GIROUX *(2016)*
Evolutionary biologist Christie Wilcox mines reams of research on venomous fauna, a vast cross-taxa group that ranges from the platypus (*Ornithorhynchus anatinus*), which delivers venom containing 83 toxins, to the Komodo dragon (*Varanus komodoensis*), whose anticoagulant-laced version bleeds victims dry. We may cringe at snakebite necrosis, but Wilcox reminds us that venoms are "complex molecule libraries" with medical potential — so safeguarding their biodiversity also preserves biochemical riches. **Barbara Kiser**

Correspondence

Reforms set to seep into India's schools

A culture of rote learning in Indian schools could be partly to blame for the "copy and paste" mentality that undermines the country's research (see A. Chaurasia *Nature* **534,** 591; 2016). Instead, children should be learning the importance of critical thinking, problem-solving and real-life application.

Attempts to abolish rote learning so far extend only to private schools (see go.nature.com/2am4jdb). However, many more children stand to gain from the innovative non-government education initiative Ekal Vidyalaya, which uses a creative educational approach through a system of one-teacher schools in rural areas and tribal villages (www.ekal.org).

Early results of public consultations by the government's Committee for Evolution of the New Education Policy and its Framework for Action promise other alternatives (see go.nature.com/2au3pej). And the 13 bold themes related to school education that have been identified as areas for improvement (see go.nature.com/2aurjby) should enable a new future.

Sanchit Misra *Banaras Hindu University, Varanasi, India.*
imsam93new@gmail.com

Kudos for female Antarctic researchers

Women scientists were prohibited from working in Antarctica until Soviet geologist Maria Klenova began her research there in 1956. Despite their contributions since, women comprise only 11% of medal winners from the Scientific Committee on Antarctic Research. Our aim is to raise the profile of influential female researchers to inspire the roughly 60% of early-career polar scientists who are women.

Notable contributions by women include the discovery of potential methane reservoirs beneath Antarctica (5 female authors out of 13: J. L. Wadham *et al. Nature* **488,** 633–637; 2012); the finding that snow melting accelerated in the twentieth century (4 of 9 authors: N. J. Abram *et al. Nature Geosci.* **6,** 404–411; 2013); and insights into life in the deep Southern Ocean (12 of 21: A. Brandt *et al. Nature* **447,** 307–311; 2007). The directors of the two largest polar institutes, the British Antarctic Survey and the Alfred Wegener Institute in Germany, are women.

To boost recognition of such achievements, we are writing referenced biographies for prominent female Antarctic scientists, and have received 170 nominations from 30 countries (see go.nature.com/2azwkjq). Examples include In-Young Ahn of the Korea Polar Research Institute, the first Asian woman to lead an Antarctic station, and Lois Jones, who in 1969 led the first all-female Antarctic research team.

Jan Strugnell* *La Trobe University, Melbourne, Australia.*
j.strugnell@latrobe.edu.au
**On behalf of 7 correspondents (see go.nature.com/2akdzbd for full list).*

Funding: would Mendel have won it?

The finding that interdisciplinary research has low funding success touches a sore spot in molecular biology (see L. Bromham *et al. Nature* **534,** 684–687; 2016). The skilful integration of physics, mathematics and biology that led to the development of molecular biology is being superseded by the use of bioinformatics tools that can process and visualize large amounts of experimental data. Yet these tools often deliver only incremental advances in complex topics (for instance, in the function of transcriptional networks).

Genuinely interdisciplinary landmark discoveries include the stochastic nature of gene expression and the realization that biological systems are 'noisy' (M. B. Elowitz *et al. Science* **297,** 1183–1186; 2002), and the finding that there was interbreeding between Neanderthals and ancestors of modern humans (R. E. Green *et al. Science* **328,** 710–722; 2010). That discovery relied on sophisticated sample-preparation methods and advanced statistical analysis to reconstruct the flow of genetic material between ancient genomes.

Historically, scientific curiosity has been driven by interdisciplinary knowledge. Gregor Mendel, for example, trained as a physicist. Modern teaching tends to gloss over the mathematical insights that his theory of inheritance required. I suspect that few biologists today could identify binomial distributions in pea-plant cross-breeding experiments and conclude that independent alleles are randomly segregated.

Daniel Hebenstreit *University of Warwick, Coventry, UK.*
d.hebenstreit@warwick.ac.uk

Funding: spot value in grant proposals

Interdisciplinary projects might have more funding success if some review-panel members had interdisciplinary research experience (see L. Bromham *et al. Nature* **534,** 684–687 (2016) and go.nature.com/2at80wd).

Such reviewers are more likely to grasp the importance of lines of investigation that fall outside disciplines. Our study on the feasibility of treating heroin users with pharmaceutical heroin, for example, called for research into whether this perceived permissiveness might influence illicit drug use and have a 'honeypot' effect (G. Bammer *Palgrave Commun.* **2,** 16017; 2016).

Interdisciplinary reviewers also recognize that disciplinary research that is not cutting-edge can still warrant funding if it sheds light on an interdisciplinary problem. Our insights into heroin-addiction treatment came from, among others, economists who determined the likely impact on the drug market; demographers who estimated the number of heroin users; and philosophers who assessed the ethics of prescribing heroin.

The grant-review process could be improved if disciplinary and interdisciplinary panel members had a better understanding of how their views interact, and if guidelines could be drawn up for their relative contributions to the overall assessment.

Gabriele Bammer *Australian National University, Acton, Australia.*
gabriele.bammer@anu.edu.au

Satellite company clarifies proposal

As chief executive of the satellite-communications company Ligado Networks, I wish to emphasize that our proposed sharing of a small block of radio frequencies with the US National Oceanic and Atmospheric Association (NOAA) will not jeopardize the delivery of weather information from satellites (see *Nature* **535,** 208–209; 2016).

We are exploring this idea through open dialogue with the US Federal Communications Commission, NOAA and the weather community. The company was invited to discuss radio frequencies at an American Meteorological Society meeting last month, and feedback on our plan's potential impact is allowing us to home in on remaining concerns and discuss solutions.

We are also exploring an alternative network for providing real-time weather data to more users at a lower cost. This would protect NOAA's existing uses of the band and expand the availability of a high-demand wireless spectrum. Our technology could deliver reliable, secure connectivity to critical industries, including those that serve public safety.

Doug Smith *Ligado Networks, Reston, Virginia, USA.*
spectrum@ligado.com

ENERGY SCIENCE

Fast track for silver

A solid composite material has been made that conducts electricity through the rapid transport of silver ions, which diffuse faster than in some liquids. The material holds promise for applications in charge–storage devices. SEE ARTICLE P.159

TOM NILGES

What happens when two compounds that have contrary properties are mixed together? Do they work against each other, or do they combine constructively to produce unexpected effects? On page 159, Chen *et al.*[1] report an impressive example of the second outcome. They have combined a material that conducts electricity purely through negatively charged electrons with one that conducts through the fast movement of positively charged ions, to create a composite that they call an artificial mixed conductor. The composite exhibits impressively fast ion diffusion, and has the potential to be of use in batteries.

The electron conductor in the composite is graphite[2], a carbon allotrope composed of layers made up of six-membered carbon rings. This layered structure means that conduction in graphite varies with the direction of the current. Graphite can conduct negatively charged electrons but can also host various highly mobile ions, and is a widely used electrode in energy-storage devices. Atomically thin layers of graphite are called graphene, and can act as a membrane that conducts protons[3].

The other component in the authors' composite is rubidium silver iodide ($RbAg_4I_5$), the best known solid conductor of silver ions at room temperature[4]. This compound can itself be thought of as a composite of silver iodide (AgI) and rubidium iodide (RbI). Silver ions are positively charged, and are large and heavy compared with most other charge carriers. But the rubidium ions in the iodide framework of $RbAg_4I_5$ are perfectly arranged to provide vacant sites for the silver ions to 'jump' into. This allows the silver ions to move almost freely in all directions through the solid.

If graphite and $RbAg_4I_5$ are combined, then a solid, mixed conductor system might be generated that enables charge transport or conduction by two different charge carriers at the interfaces between the two compounds, potentially offering extremely high conductivity. Conductors that naturally allow conduction through both electron and ion transport have been widely studied and are used in processes that benefit from such optimized conductivity. For example, they have applications in energy-storage devices such as batteries and supercapacitors, and in sensor devices[5–7].

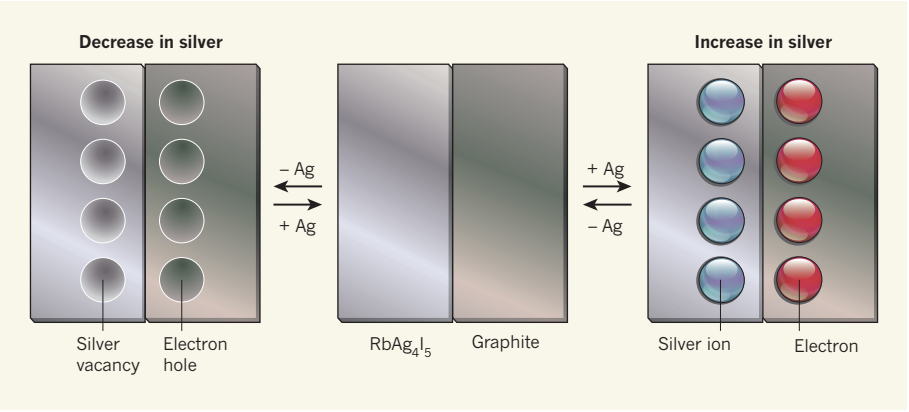


Figure 1 | Interfacial mass transfer. Chen *et al.*[1] have prepared a composite of rubidium silver iodide ($RbAg_4I_5$, a material that conducts using silver ions as charge carriers) and graphite (in which electrons carry charge). When the composite is connected to a silver electrode (not shown), the authors observe reversible rapid movement of silver ions that leads to either a reduction or an increase in the amount of silver in the $RbAg_4I_5$, depending on the direction of the current. When the amount of silver decreases, electron holes (quasiparticles caused by the absence of electrons) in the graphite compensate for vacancies caused by the absence of silver ions in the $RbAg_4I_5$, at the interfaces between particles of the two materials within the composite. When silver is added, the extra silver ions in the $RbAg_4I_5$ are balanced by electrons in the graphite.

Duality of charge transport has another advantage: it can be used to change the stoichiometry (the ratio of atom types described by a chemical formula) of a conductor, because the addition or removal of a given charge carrier can be compensated for by adding or removing the oppositely charged carriers. This enables fast transport, storage and redistribution of mass, which is also useful for charge-storage devices. For example, an extra positively charged ion such as Ag^+ can be compensated for by the addition of a negatively charged electron, whereas a vacant Ag^+ ion can be balanced by adding an electron hole (a quasiparticle corresponding to the absence of an electron; Fig. 1). In pure electron- or ion-conducting systems, such compensation processes are disfavoured or almost impossible, in most cases because of the lack of oppositely charged carriers. A solution to this problem that allows certain stoichiometry changes could be to combine both types of system to form a hybrid.

Chen and colleagues' graphite–$RbAg_4I_5$ composite is just such a clever combination. The authors prepared the material by grinding the components together in a mortar and then melting the mixture, to bring the two types of conductor into intimate contact with each other. In a series of experiments, the authors observed impressively fast (occurring within seconds) and pronounced stoichiometry changes (from approximately -10^{-5} to about 4×10^{-6} for silver) at the interfaces between the two compounds. This behaviour was combined with an extraordinarily high diffusion of silver ions at room temperature.

But how do the components of the composite enable stoichiometric changes? In the case of silver being added to the composite, the extra silver is stored within the ion conductor ($RbAg_4I_5$), whereas the compensatory electrons are hosted by the electron conductor (graphite). Both compounds contribute their best talents to this joint effort — $RbAg_4I_5$ effectively transports silver ions and stores them at vacant surface sites, and graphite acts as an electron sponge. The situation is different in the case of silver being removed from the composite: $RbAg_4I_5$ releases silver and forms vacancies that are compensated for by electron holes in the graphite. Once again, the job is shared by the two compounds.

Chen and co-workers also provide a detailed analysis of the physics behind the observed fast

diffusion process, and show that the classical theory of chemical diffusion must be reconsidered in the case of interface-driven job-sharing processes. In particular, there should be a reassessment of the roles of chemical capacitance (a material's ability to take up or release chemical components such as silver ions) and of electrostatic energy in changing the charge-carrier concentration at the interface.

Finally, the authors built two all-solid-state prototype energy-storage devices — a battery and a supercapacitor — to demonstrate potential practical applications of their composite. The battery can be reversibly charged and discharged using extremely high currents (and therefore within 0.05 seconds), whereas the supercapacitor provides ultrafast charge release, which is needed for various applications of these devices. Both effects are a direct result of the high mass transport in the system and its compositional flexibility.

It will be exciting to see whether the concept of an artificial mixed conductor system can be transferred to other solid ion conductors, such as the promising 'argyrodite-type' solids, in which lithium ions have unusually high mobility[8,9]. Another question is whether graphite is the optimal electron conductor in these mixed conductors. Perhaps graphene sheets, or graphite consisting of just a few stacked graphene sheets, could be used instead, to optimize the number of interfaces per unit volume between the two types of conductor. Materials scientists, chemists and physicists will no doubt be keen to adopt this concept to create materials that have hybrid functionality, potentially opening up fresh applications. ■

Tom Nilges *is in the Department of Chemistry, Technical University of Munich, 85748 Garching bei München, Germany.*
e-mail: tom.nilges@lrz.tum.de

1. Chen, C.-C., Fu, L. & Maier, J. *Nature* **536,** 159–164 (2016).
2. Bernal, J. D. *Proc. R. Soc. A* **106,** 749–773 (1924).
3. Hu, S. *et al. Nature* **516,** 227–230 (2014).
4. Geller, S. *Science* **157,** 310–312 (1967).
5. Maier, J. *Ann. Phys.* **15,** 469–479 (2006).
6. Riess, I. *Solid State Ionics* **157,** 1–17 (2003).
7. Rivnay, J. *et al. Nature Commun.* **7,** 11287 (2016).
8. Deiseroth, H.-J. *et al. Angew. Chem. Int. Edn* **47,** 755–758 (2008).
9. Deiseroth, H.-J. *et al. Z. Anorg. Allg. Chem.* **637,** 1287–1294 (2011).

NEUROSCIENCE

Nanocolumns at the heart of the synapse

A nanocolumn spans the synaptic cleft between neurons, connecting regions of neurotransmitter molecule release and capture. This discovery informs on mechanisms of synaptic organization and regulation. SEE LETTER P.210

STEPHAN J. SIGRIST & ASTRID G. PETZOLDT

The sophisticated human brain forms the foundation of all the cognitive processes that define us as self-conscious and social individuals. These processes are fundamentally based on the operation of a single functional unit — the synapse, which enables rapid signal transmission between neurons. Synapses are composed of two small, highly specialized compartments, one on the presynaptic (transmitting) side and one on the postsynaptic (receiving) side of the small gap that separates the two neurons. Structures spanning this synaptic cleft to coordinate these compartments have been suggested[1], but direct evidence for their existence remains scarce. On page 210, Tang *et al.*[2] use a combination of elaborate super-resolution light microscopy and mathematical modelling to provide evidence for the existence of discrete, protein-based nanocolumns that connect the pre- and postsynaptic compartments.

During neuronal signalling, electrical impulses called action potentials trigger the release of neurotransmitter molecules from the presynaptic neuron. Release involves fusion of neurotransmitter-containing synaptic vesicles with a region of the cell membrane called the active zone, which faces the synaptic cleft. Vesicle docking and fusion does not occur in isolation, but within an extended protein scaffold made up of several large multi-domain proteins[3] that provides sites for synaptic-vesicle fusion.

Dissecting the organizational principles of these scaffolds was, for many years, achievable only by electron microscopy, which is not compatible with live imaging. However, this limitation has been overcome, thanks to the development of super-resolution light-microscopy techniques[4], which allow the efficient visualization of distinct protein architectures. One such study[5] has revealed that presynaptic scaffolds physically contact synaptic vesicles, perhaps promoting their docking and priming for neurotransmitter release at defined fusion sites. Other studies have shown that one scaffold protein, RIM, has a prominent role in synaptic-vesicle docking — RIM interacts with proteins of the MUNC-13 family[6,7] to promote clustering of calcium-channel proteins[8], which in turn trigger fusion processes.

In a quest to further decipher the nano-architecture of presynaptic active zones, Tang *et al.* turned to a high-resolution form

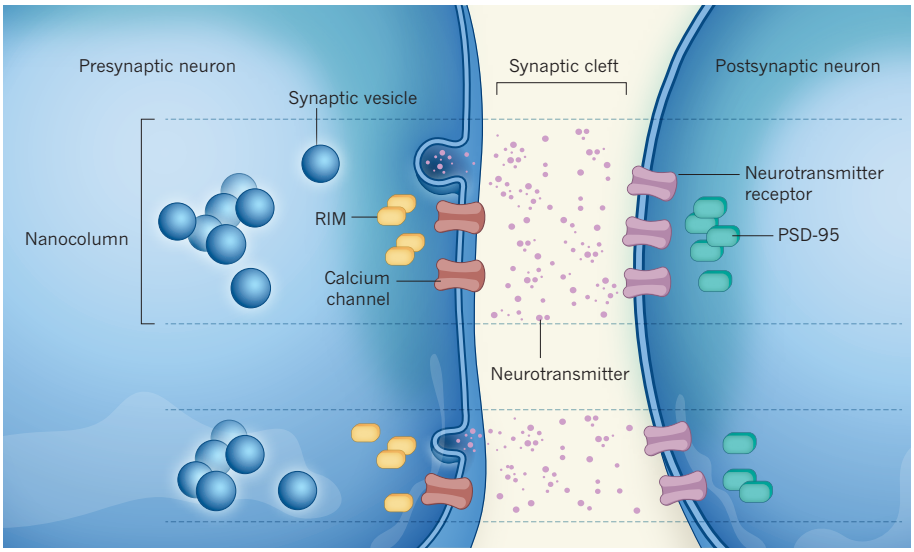


Figure 1 | Architecture of a synapse. Tang *et al.*[2] report that the synaptic connections between neurons are bridged by nanocolumn structures. The scaffold protein RIM is enriched in 80-nanometre-wide clusters at sites on the presynaptic membrane to which synaptic vesicles fuse close to calcium channel proteins and release neurotransmitter molecules into the synapse. On the postsynaptic neuron, sites rich in the scaffold protein PSD-95 contain clusters of neurotransmitter receptor proteins. RIM-rich and PSD-95-rich regions align to define the nanocolumn.

of light microscopy called stochastic optical reconstruction microscopy (STORM)[9]. The authors used 3D STORM to study synapses between *in-vitro* cultured mouse neurons derived from the brain's hippocampus region, which is involved in learning and memory. The synapses under observation release the neurotransmitter glutamate. This analysis revealed that RIM is confined to protein nanoclusters of around 80 nanometres in diameter that lie close to the active zone. By contrast, other scaffold proteins and fusion factors, such as MUNC-13 and Bassoon, showed a more uniform distribution.

Is the position of these RIM-rich nanoclusters related to vesicle-fusion sites? The researchers monitored fusion events at the presynaptic membrane using a protein-based sensor that fluoresces following vesicle–membrane fusion. Mathematical modelling of the fluorescence patterns revealed that fusion sites are restricted to particular regions of the membrane. Moreover, a different form of super-resolution light microscopy called photoactivated localization microscopy that allows live imaging, confirmed that RIM density increased within 40 nm of these fusion sites.

The authors next investigated whether the RIM-rich fusion sites might be coordinated with the position of the postsynaptic apparatus dedicated to receiving the neurotransmitter signal. A sophisticated scaffold resides close to the postsynaptic membrane — a part of which is the multi-domain protein PSD-95, which is involved in the clustering of AMPA- and NMDA-type glutamate receptor proteins[10,11]. Precise measurement of RIM and PSD-95 densities revealed a clear spatial correlation between the components. Tang *et al.* therefore concluded that a nanoscale columnar structure spans the synaptic cleft, bringing RIM-enriched sites of synaptic-vesicle fusion face-to-face with postsynaptic PSD-95 nanodomains (Fig. 1).

Finally, Tang and colleagues asked if the nanocolumns could be a stable architectural motif or whether they are involved in the regulatory changes in synaptic strength that are crucial for cognitive functions. The authors pharmacologically activated NMDA receptors to depress synaptic strength. Although there was no immediate change in the architecture of the nanocolumn, after 25 minutes a subset of RIM nanoclusters suddenly grew larger — notably only those lying opposite PSD-95 nanodomains and residing in nanocolumns. Thus, retrograde signals that mediate the upregulation of presynaptic release in response to postsynaptic changes might specifically target the scaffold proteins and release machinery located opposite the postsynaptic glutamate receptors to modulate synaptic strengthening. As such, the nanocolumn could provide an important regulatory platform.

This study generates pressing questions. For instance, to understand the physical nature of the nanocolumns, it would be interesting to determine what regulates their formation. Trans-synaptic pairs of cell-adhesion membrane proteins are obvious candidates for mediating nanocolumn formation. Perhaps such adhesion molecules ultimately control the positioning and recruitment of RIM.

Alternatively, diffusible signals might cross the cleft and specifically trigger assembly of nanocolumns on the scale of a few tens of nanometres. In addition, RIM itself could be involved in nanocolumn formation — RIM contains a central domain that binds to the intracellular part of calcium channels[12], which ultimately trigger synaptic-vesicle fusion.

In the future, the nanocolumn concept should be validated and extended by investigating more proteins, including synaptic cell-adhesion proteins and other cytoplasmic scaffold proteins, and by combining imaging with genetic manipulation. Although the details of trans-synaptic coordination and the proteins involved might turn out to vary between synapse types and organisms, the nanocolumnar architectural motif could be a fundamental and generic building principle for synapses. ■

Stephan J. Sigrist *and* **Astrid G. Petzoldt**
are at the Institute of Biology, Free University of Berlin, 14195 Berlin, Germany, and the Cluster of Excellence NeuroCure, Charité–Universitätsmedizin Berlin.
e-mails: stephan.sigrist@fu-berlin.de; astrid.petzoldt@fu-berlin.de

1. Chia, P. H., Li, P. & Shen, K. *J. Cell Biol.* **203,** 11–22 (2013).
2. Tang, A.-H. *et al. Nature* **536,** 210–214 (2016).
3. Ackermann, F., Waites, C. L. & Garner, C. C. *EMBO Rep.* **16,** 923–938 (2015).
4. Maglione, M. & Sigrist, S. J. *Nature Neurosci.* **16,** 790–797 (2013).
5. Matkovic, T. *et al. J. Cell Biol.* **202,** 667–683 (2013).
6. Fernández-Busnadiego, R. *et al. J. Cell Biol.* **201,** 725–740 (2013).
7. Deng, L., Kaeser, P. S., Xu, W. & Südhof, T. C. *Neuron* **69,** 317–331 (2011).
8. Han, Y., Kaeser, P. S., Südhof, T. C. & Schneggenburger, R. *Neuron* **69,** 304–316 (2011).
9. Rust, M. J., Bates, M. & Zhuang, X. *Nature Meth.* **3,** 793–795 (2006).
10. Nair, D. *et al. J. Neurosci.* **33,** 13204–13224 (2013).
11. MacGillavry, H. D., Song, Y., Raghavachari, S. & Blanpied, T. A. *Neuron* **78,** 615–622 (2013).
12. Kaeser, P. S. *et al. Cell* **144,** 282–295 (2011).

This article was published online on 27 July 2016.

SYSTEMS NEUROSCIENCE

A modern map of the human cerebral cortex

An authoritative map of the modules that make up the cerebral cortex of the human brain promises to act as a springboard for greater understanding of brain function and disease. SEE ARTICLE P.171

B. T. THOMAS YEO & SIMON B. EICKHOFF

The human brain's cerebral cortex is crucial for sensory and motor processing, as well as for mental functions such as interpreting language and logical reasoning, the complexity of which distinguishes us from other animals. On page 171, Glasser *et al.*[1] describe an updated map of the human cerebral cortex. This long-awaited advance provides a reference atlas that will allow those researching brain structure, function and connectivity to work within a common, systems-neuroscience framework.

Regional differentiation within the cerebral cortex has long prompted attempts to identify the cortex's distinct compartments, from classical neuroanatomical studies at the beginning of the twentieth century[2] to modern non-invasive, *in vivo* methods based on magnetic resonance imaging (MRI). Such endeavours are complicated by the fact that every location in the brain can be described by an almost infinite set of features, including density of receptor proteins for various neurotransmitter molecules, long-range connections to other parts of the brain, and specialization for neural computations that support specific functions. Almost all previous studies have attempted to delineate cortical compartments using a single feature (Fig. 1). By contrast, Glasser and colleagues capitalize on the unprecedented quality and breadth of MRI data gathered by the Human Connectome Project, the aim of which is to elucidate the neural pathways that underlie brain function and behaviour using cutting-edge brain-imaging methods[3].

MRI provides unparalleled access to the living brain. A single MRI machine can take many different measurements (known as modalities) — from establishing the relative density of neuron-insulating myelin sheaths to determining the thickness of the cortex, both of which can vary sharply between

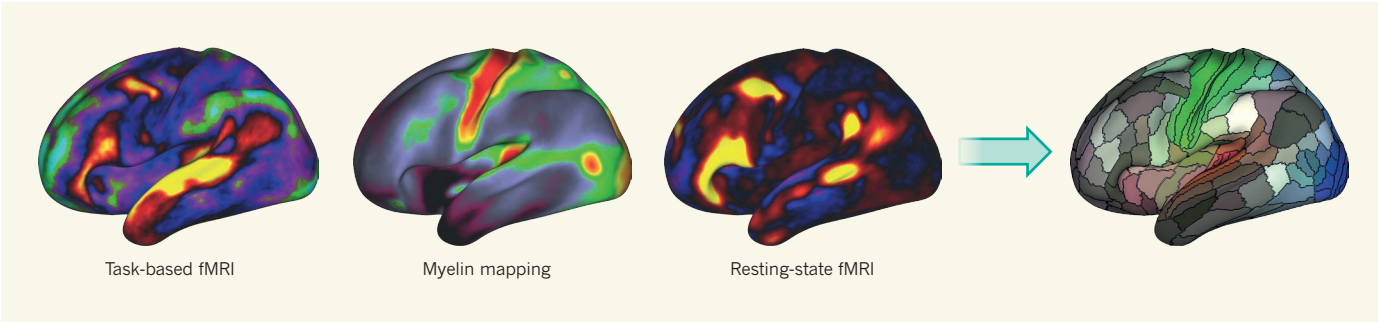


Figure 1 | Mapping function in the brain. Glasser *et al.*[1] defined distinct regions in the human cerebral cortex using a combination of brain-mapping techniques that have previously been used only separately, including: task-based functional magnetic resonance imaging (fMRI), which informs on the functions of different regions; relative density of the neuron-sheathing substance myelin, which provides information about cortical architecture; and resting-state fMRI, which details neural connectivity within and between different regions. In each of these three panels, colours provide a heat map of the measurements. The result is a map that delineates 360 distinct cortical areas. Different colours represent how connected each area is to sensory inputs (hearing, red; touch, green; vision, blue) and to systems involved in cognition (light and dark). Mixed colours show areas in which functional systems overlap. (Images taken from ref. 1.)

cortical areas. Furthermore, functional MRI (fMRI) can measure the changes in blood flow that accompany mental tasks, as well as whole-brain activity in resting states, providing information about regional neural activity that accompanies different brain states. The authors' integration of information from several MRI modalities not only moves this work closer than previous attempts to the classical definition of a cortical area, but also has several key advantages over other investigations.

First, some modalities reveal borders not clearly reflected in others. For instance, the border between areas 3a and 3b of the somatosensory cortex (which processes information about touch and pain) is easily delineated by myelin mapping, but not by resting-state fMRI. As another example, Glasser *et al.* developed a resting-state fMRI technique that maps topographic neural connectivity within the visual cortex. The sharp transition between levels of topographic connectivity across area boundaries allows much clearer delineation of discrete areas involved in early stages of visual processing than do myelin maps or conventional resting-state fMRI approaches[4,5].

Second, convergence across different MRI modalities reduces the likelihood of misdefining borders as a result of feature-specific noise or bias. This is important, given the indirect nature of most modalities — for example, fMRI measures the blood-flow changes that accompany neuronal activity as a proxy for neuronal activity itself. Consequently, complex computational pre-processing is often necessary to differentiate signal from noise. Agreement across modalities increases confidence that borders reflect biological reality rather than measurement biases.

Finally, an integrative approach better equips researchers to describe the properties of each area, as exemplified by Glasser and colleagues' comprehensive supplementary material. The authors find, for instance, that a cortical area characterized in the 1950s by its low myelin content[6] seems to be involved in language processing as measured by task-based fMRI — a finding consistent with a recent meta-analysis of more than 10,000 imaging experiments across 83 behavioural tasks[7]. Therefore, Glasser and colleagues' map represents the convergence of decades of classical neuroanatomical studies with modern non-invasive studies.

In contrast to the burgeoning field of resting-state fMRI mapping, which has largely focused on fully automatic approaches to divide the brain into parcels that have homogeneous connectivity patterns[8], Glasser and colleagues used a semi-automatic approach that explicitly incorporates prior knowledge from neuroanatomical studies to define the borders in their map. This inclusion represents a crucial and long overdue advance over agnostic, exclusively computational approaches. However, using prior knowledge to choose which modalities to trust in cases of conflicting evidence entails the danger of introducing confirmatory biases. Moreover, it could result in differential mapping quality between areas in which there is relevant, well-known information — such as the somatosensory and visual cortices — and those for which less knowledge exists, such as the prefrontal and parietal cortices. The latter pair is of particular interest to many neuroscientists, because these areas compute most functions that are specific to humans. Indeed, given that the authors explicitly ignore certain modality information for their data set that is functionally meaningful but fractionates classical cortical areas, further investigation will be crucial to understand how borders that are strongly demarcated in only one modality can be differentiated from modality-specific noise.

On a related theme, although Glasser *et al.* have delineated 360 cortical areas, these regions could potentially be subdivided into smaller, more-uniform units that are less distinct from each other. For example, different portions of the somatosensory cortex that represent distinct body parts might be considered as distinct computational units. Furthermore, examples of new areas being defined with the advent of more-sensitive or complementary methods are commonplace[9]. As such, it remains unclear what the 'optimal' number of areas to be defined is — let alone the 'correct' number. We suspect that the optimal number might be application-dependent. The authors' work, although seminal, will therefore probably not be the final word on this topic.

A key innovation in the current study is an automatic algorithm that seeks to delineate cortical areas in individual human subjects, a much more complex task than producing a map of the average brain. Previous work has attempted to estimate, in individual subjects, 10–20 functional networks (for example, see ref. 10), but Glasser and colleagues' goal of delineating 360 areas is more ambitious. Capturing inter-individual biological variability and differentiating such variability from measurement noise is essential to understand the relationship between brain organization and individual differences in behaviour, as well as for clinical applications.

The authors' validation of this algorithm focused on only a small portion of the cortex, so further investigation will be crucial. Nevertheless, their work represents a major step towards individual-specific 'biomarkers' of brain dysfunction, because individual-specific quantities of each area, such as grey-matter volume or connectional strength to other areas, can now be computed, and could be strongly predictive of individual differences in behaviour or disease.

Glasser and co-workers' atlas is the first multimodal map targeted at defining cortical areas, and therefore represents a major advance in human brain mapping. It is now up to researchers to use the anatomical framework provided, compare it with alternative approaches to mapping the human brain, and populate the defined areas with functional

and disease-related information. By doing so, we can begin to integrate multimodal data to understand how individual differences in brain organization can explain differences in function, behaviour and disorder. ■

B. T. Thomas Yeo *is in the Department of Electrical & Computer Engineering, Clinical Imaging Research Centre, Singapore Institute for Neurotechnology & Memory Network Programme, National University of Singapore, 117456 Singapore.* **Simon B. Eickhoff** *is in the Department of Clinical Neuroscience and Medical Psychology, Heinrich-Heine University, 40225 Düsseldorf, Germany, and the Institute for Neuroscience and Medicine (INM-1), Jülich, Germany.*
e-mails: thomas.yeo@nus.edu.sg; s.eickhoff@fz-juelich.de

1. Glasser, M. F. *et al. Nature* **536,** 171–178 (2016).
2. Brodmann, K. *Vergleichende Lokalisationslehre der Großhirnrinde: in ihren Prinzipien dargestellt auf Grund des Zellenbaues* (J. A. Barth, 1909); English transl. available in Garey, L. J. *Brodmann's Localization in the Cerebral Cortex* (Smith Gordon, 1994).
3. WU-Minn HCP Consortium. *NeuroImage* **80,** 62–79 (2013).
4. Buckner, R. L., Krienen, F. M. & Yeo, B. T. T. *Nature Neurosci.* **16,** 832–837 (2013).
5. Gordon, E. M. *et al. Cereb. Cortex* **26,** 288–303 (2014).
6. Hopf, A. *J. Hirnforsch.* **2,** 311–333 (1956).
7. Yeo, B. T. T. *et al. Cereb. Cortex* **25,** 3654–3672 (2015).
8. Eickhoff, S. B., Thirion, B., Varoquaux, G. & Bzdok, D. *Hum. Brain Mapp.* **36,** 4771–4792 (2015).
9. Amunts, K. *PLoS Biol.* **8,** e1000489 (2010).
10. Wang, D. *et al. Nature Neurosci.* **18,** 1853–1860 (2015).

This article was published online on 20 July 2016.

CANCER

Endothelial-cell killing promotes metastasis

To migrate into the lungs, cancer cells in the bloodstream must cross the lung's endothelial-cell barrier. A study shows that cancer cells can achieve this feat by signalling to induce endothelial-cell death. SEE LETTER P.215

CLAUDIO R. ALARCÓN & SOHAIL F. TAVAZOIE

Cancer cells often migrate from where the cancer initially formed, to colonize other parts of the body in a process called metastasis, which is associated with poor clinical prognosis. On page 215, Strilic *et al.*[1] uncover a surprising mechanism that migrating cancer cells in the bloodstream use to cross the lung's barrier of endothelial cells. The authors show that cancer cells send a signal that makes endothelial cells undergo a type of cell-death program called necroptosis (also known as programmed necrosis). Once in the lung, the cancer cells form lethal metastatic colonies.

The past three decades have provided increasing evidence that supports the key roles of endothelial cells in the formation and progression of tumours to a malignant state that has a poor prognosis for the patient[2–4]. Tumour cells rely heavily on the endothelial cells of blood vessels to enable continued tumour growth, because tumours need blood vessels to obtain oxygen and nutrients and expel metabolic waste.

Tumour cells exploit and manipulate endothelial cells by using intricate signalling mechanisms, such as those involving protein factors, secreted by tumours, that attract and remodel endothelial cells. Remodelling of blood vessels by tumour-derived proteins can enable cancer cells that reside in the primary site of tumour growth to enter the blood circulation, providing an escape route for the cells to reach distant organs[5,6]. After entering the bloodstream, cancer cells must cross the endothelial barriers that prevent them from entering other organs (Fig. 1a). In certain tissues, such as the lung or brain, the interface between the tissue and the bloodstream is relatively impenetrable to tumour cells[6].

Strilic and colleagues' work began with an observation made when tumour cells and endothelial cells were cultured together *in vitro*. The researchers noted that such co-culture leads to an increase in endothelial-cell death. However, rather than exhibiting the typical cell-shape changes and molecular features of apoptosis, the most common form of programmed cell death, the dying endothelial cells exhibited features associated with another cell-death program called necroptosis. For example, the dying cells exhibited compromised cell-membrane integrity, as monitored by dye uptake.

To confirm the observed cell-death mechanism, the authors inhibited proteins that mediate necroptosis[7], and found that this inhibited tumour-induced endothelial-cell death, whereas perturbing apoptotic signalling did not. They found that this necroptotic cell-death program was activated in both human and mouse endothelial cells exposed to a wide variety of cancer cell lines. Moreover, intravenous injection of mouse melanoma skin-cancer or lung-cancer cells into mice caused lung endothelial cells to undergo necroptotic death.

What advantage does killing endothelial cells afford tumour cells? Strilic and colleagues carried out *in vitro* experiments in which they inhibited necroptosis and observed reduced tumour-cell migration across an endothelial-cell monolayer, leading the authors to propose

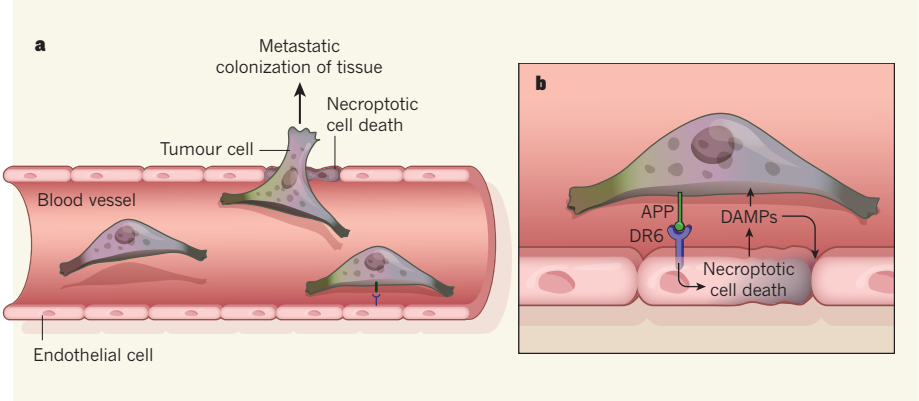


Figure 1 | Tumour cells migrate into tissues by killing cells that block their entry. **a**, Strilic *et al.*[1] show that tumour cells can induce necroptotic cell death of blood-vessel endothelial cells, which enables migrating (metastatic) cancer cells in the bloodstream to cross the endothelial-cell barrier and enter the adjacent tissue to colonize a new tumour site. **b**, Necroptotic endothelial-cell death is induced by amyloid precursor protein (APP) on the tumour surface, which interacts with the death receptor 6 protein (DR6) on endothelial cells. Tumour-cell migration from the bloodstream into the adjacent tissue may be enhanced either directly, as a consequence of endothelial-cell death and the resulting disruption of the endothelial barrier, or indirectly because of the release of damage-associated molecular pattern molecules (DAMPs) from dying endothelial cells that could open the endothelial barrier between cells or enhance tumour migration properties.

that tumour-induced necroptosis enhanced tumour-cell migration across the endothelial barrier. The authors made similar findings in *in vivo* experiments using genetic inactivation of the RIPK3 kinase enzyme, a necroptosis regulator, in endothelial cells. Inactivation of RIPK3 prevented endothelial-cell death, and reduced the ability of cancer cells to cross the endothelial barrier and enter the lung. Metastastic tumour-colony formation was reduced upon genetic or pharmacological inhibition of endothelial-cell necroptosis, indicating that tumour-induced endothelial-cell killing exerted control over metastasis.

How does endothelial-cell death enhance tumour-cell migration across the endothelial barrier? Strilic and colleagues propose various mechanisms. Tumour cells could migrate through gaps left in the endothelial barrier by dead endothelial cells. Another possibility is that damage-associated molecular pattern molecules (DAMPs), such as ATP released from necroptotic endothelial cells, could act on neighbouring endothelial cells to open the endothelial barrier by enabling tumour-cell migration between neighbouring endothelial cells that are usually bound together to form an impermeable barrier, and/or these signals could act directly on tumour cells to enhance their migration across the barrier[8].

How do tumour cells induce endothelial necroptosis? The authors used a combination of molecular, pharmacological and genetic approaches to show that amyloid precursor protein (APP) on the surface of tumour cells induces necroptotic cell death by interacting with death receptor 6 (DR6) on endothelial cells (Fig. 1b). Consistent with this, pharmacological inhibition of DR6 signalling — achieved by injecting mice with a 'decoy' version of the DR6 receptor — inhibited metastasis.

Strilic *et al.* provide compelling evidence to support the existence of intricate signalling interactions between migrating tumour cells in the bloodstream and the blood-vessel endothelium that promote tumour-cell metastatic migration into tissue and subsequent tissue colonization. These findings raise a series of intriguing issues. Only a small fraction of endothelial cells cultured *in vitro* with tumour cells are induced to undergo necroptotic death. Discovering the molecular determinant that governs which endothelial cells die is a key challenge. The authors reveal that only approximately 10% of endothelial cells express DR6 and are thus susceptible to APP-mediated cell death.

It will be important to understand the mechanisms that regulate which fraction of endothelial cells express DR6, and whether cancer cells can regulate the susceptibility of endothelial cells to necroptosis by modulating DR6 expression on the cells. Microscopy analysis of human tumours could be used to reveal whether an increased fraction of DR6-expressing endothelial cells is associated with the propensity for lung metastatic progression. Perhaps molecular signals from the endothelium to tumour cells regulate expression or cleavage of APP on tumour cells — thus having an effect on endothelial-cell necroptosis. Such endothelial-cell-derived signals have roles in epithelial-cell fate and function[9].

In addition to the mechanisms proposed by the authors, another mechanism by which endothelial-cell necroptosis might enhance tumour migration into tissue could be mediated by ATP. Release of ATP from dying endothelial cells might promote the survival of tumour cells during their migration through the endothelial barrier into the tissue[10] — a process that can cause traumatic tumour-cell deformation and death. Live-cell microscopy imaging of tumour- and endothelial-cell dynamics during this interaction[11] may be an ideal means of determining which of the intriguing potential cellular mechanisms proposed by the authors might underlie tumour-cell migration across the endothelial barrier. ■

Claudio R. Alarcón *and* **Sohail F. Tavazoie**
are at the Laboratory of Systems Cancer Biology, The Rockefeller University, New York, New York 10065, USA.
e-mails: calarcon@rockefeller.edu; stavazoie@rockefeller.edu

1. Strilic, B. *et al. Nature* **536,** 215–218 (2016).
2. Folkman, J. *Cancer Res.* **46,** 467–473 (1986).
3. Dvorak, H. F. *Cancer J.* **21,** 237–243 (2015).
4. Ferrara, N. *Arterioscler. Thromb. Vasc. Biol.* **29,** 789–791 (2009).
5. Minn, A. J. *et al. Nature* **436,** 518–524 (2005).
6. Gupta, G. P. *et al. Nature* **446,** 765–770 (2007).
7. Pasparakis, M. & Vandenabeele, P. *Nature* **517,** 311–320 (2015).
8. Schumacher, D., Strilic, B., Sivaraj, K. K., Wettschureck, N. & Offermanns, S. *Cancer Cell* **24,** 130–137 (2013).
9. Ding, B. S. *et al. Nature* **505,** 97–102 (2014).
10. Furlow, P. W. *et al. Nature Cell Biol.* **17,** 943–952 (2015).
11. Headley, M. B. *et al. Nature* **531,** 513–517 (2016).

This article was published online on 3 August 2016.

CELL BIOLOGY

The TORC1 pathway to protein destruction

A study of the proteasome — a protein-degradation complex — reveals an evolutionarily conserved pathway that acts through the protein kinase TORC1 to adjust proteasome levels in response to cellular needs. SEE ARTICLE P.184

LYNNE CHANTRANUPONG & DAVID M. SABATINI

To maintain amino-acid and protein levels, cells must couple nutrient availability to protein synthesis and turnover. Central to this process is the enzyme called target of rapamycin complex 1 (TORC1) kinase, a master growth controller that integrates diverse environmental inputs to coordinate many metabolic processes[1]. Rousseau and Bertolotti[2] reveal on page 184 that inhibition of TORC1 increases levels of the proteasome — a large protein complex involved in cellular protein degradation — to promote cell survival under stressful conditions. Consistent with previous reports[3–5], the new work identifies TORC1 as a central regulator of proteasome homeostasis. However, the relationship between TORC1 and the control of proteasome function seems to be complex, because TORC1 can regulate the proteasome through multiple mechanisms that depend on the particular cellular context[3–5].

The proteasome functions in one of the main protein-degradation pathways in cells, the ubiquitin–proteasome system[6]. In this pathway, a multi-enzymatic cascade covalently links the small polypeptide ubiquitin to proteins. This modification is recognized by the proteasome, which degrades ubiquitinated proteins to produce peptide mixtures that can then replenish the intracellular pool of amino acids[6].

The proteasome comprises a multisubunit core particle, which carries out protein degradation, and up to two additional regulatory particle components that facilitate substrate recognition, removal of ubiquitin, and protein unfolding and translocation into the proteasome[6]. Inhibition of the proteasome results in a lethal shortage of amino acids[7]; therefore cells must maintain adequate proteasome levels to survive. However, the mechanisms that govern the assembly and regulation of this complex molecular machine, particularly under stressful conditions, are not fully understood.

The discovery[8] in yeast of Adc17, a stress-induced regulatory particle assembly chaperone protein (RAC), offers an insight into the mechanism of proteasome regulation.

Rousseau and Bertolotti used Adc17 as a starting point to investigate the proteasome. They treated yeast with the antibiotic tunicamycin to induce the unfolded-protein response, a cellular stress response to the presence of misfolded or unfolded proteins. They found that yeast upregulates Adc17 levels in the presence of tunicamycin, and that loss of the protein Sfp1, a negative regulator of TORC1, abrogates this increase of Adc17. The authors established that the increase in Adc17 requires inhibition of TORC1. Pharmacological suppression of TORC1 by the compound rapamycin or genetic inhibition of TORC1 by inactivation of *KOG1*, which encodes an essential TORC1 subunit, are sufficient to increase the expression not only of Adc17, but also of all other known RACs and of proteasome subunits.

To understand how TORC1 might mediate an increase in proteasome abundance, the authors focused on Mpk1, a yeast enzyme known as a mitogen-activated protein kinase (MAPK) that functions downstream of TORC1, and which is essential for the survival of cells in which tunicamycin has induced a stressful increase of unfolded proteins. Rousseau and Bertolotti found that Mpk1 is required for the TORC1-mediated increase in RACs and proteasome subunits (Fig. 1a). Neither the abundance of their messenger RNAs nor the protein stability of these RACs and proteasome subunits was altered in response to rapamycin-mediated inhibition of TORC1, which indicates that the increased levels of these proteins probably occur though regulation of mRNA translation.

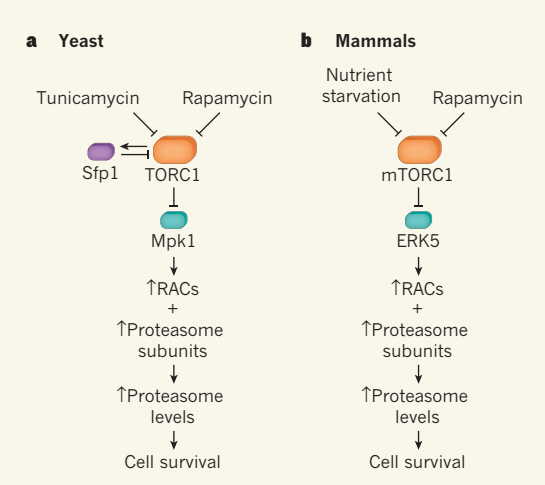


Figure 1 | Evolutionarily conserved regulation of proteasome abundance. **a**, Rousseau and Bertolotti[2] report that activation of the yeast mitogen-activated protein kinase enzyme (MAPK) known as Mpk1 mediates an increase in the levels of regulatory particle assembly chaperone proteins (RACs) and subunits required for the formation of the proteasome complexes that mediate protein degradation. Mpk1 is activated by inhibition of TORC1 protein kinase activity. TORC1 can be inhibited by tunicamycin or rapamycin treatment or by the action of the protein Sfp1. **b**, The authors also show that this process of proteasomal regulation is conserved in mammals — ERK5, the mammalian protein kinase most similar to Mpk1, is required to upregulate mammalian proteasome levels through an increase in RACs and proteasome subunits upon mTORC1 inhibition by compounds such as rapamycin or by nutrient starvation. In both yeast and mammals, an increase in proteasome abundance is necessary for cell survival under stress.

An enhanced proteasomal capacity enables cells to adapt to the rising demand for protein degradation that accompanies stress. The absence of proteasome induction, as tested by Rousseau and Bertolotti using cells in which the gene for Mpk1 had been deleted, severely impairs the clearance of ubiquitinated proteins and of well-characterized reporter substrates used to monitor proteasomal activity.

The authors found that in mammalian cells, ERK5, the mammalian equivalent of Mpk1, also facilitates a rapid rise in RAC and proteasome levels when mTORC1 (the mammalian equivalent of TORC1) is inhibited (Fig. 1b). Thus, the TORC1 and Mpk1 pathway is an evolutionarily conserved regulator of proteasomal homeostasis.

Rousseau and Bertolotti's work contributes an additional perspective to the current debate about the exact relationship between TORC1/mTORC1 and the regulation of proteasome function. Consistent with the model proposed by Rousseau and Bertolotti, a study by Zhao *et al.*[4] found that acute pharmacological inhibition of mTORC1 in the HEK293 mammalian cell line upregulates protein degradation by the proteasome. However, a report by Zhang *et al.*[3] reveals nuances in the regulation of the proteasome by mTORC1, and finds that in the absence of the protein TSC2, a major inhibitor of the mTORC1 pathway, the transcription factor NRF1 mediates an increase in levels of the proteasome and of intracellular amino acids.

The differences between these three studies[2–4] probably arise from variations in the extent to which mTORC1 is perturbed. Under acute mTORC1 inhibition, as studied by Rousseau and Bertolotti[2] and Zhao *et al.*[4], upregulation of the proteasome would increase amino-acid pools and permit the translation of proteins necessary for survival. mTORC1 inhibition induces autophagy, another major intracellular protein-degradation pathway that removes proteins in bulk from the cytoplasm and delivers them to an organelle called the lysosome for breakdown[1]. In combination, the rapid and coordinated activation of both the autophagic and proteasomal arms of protein degradation would be beneficial to cells as a mechanism to increase amino-acid levels under stress or nutrient deprivation.

However, under states of prolonged mTORC1 hyperactivation — for example, when TSC2 is lost, as investigated by Zhang *et al.*[3] — cells may also need to increase proteasomal capacity to counteract unrestrained consumption of resources driven by sustained mTORC1 activity. It will be informative to compare the regulation of the proteasome in genetic models in which mTORC1 is constitutively active but not hyperactivated — for example, in mice that have a constitutively active Rag GTPase enzyme[9].

How TORC1 inhibition increases proteasome-dependent degradation is another question requiring further investigation. Rousseau and Bertolotti found that this upregulated proteolysis depends on elevated proteasome levels, whereas the study by Zhao *et al.*[4] found that enhanced ubiquitination drives protein breakdown without a change in proteasome content or activity. It will also be of interest to determine whether specific proteins are preferentially targeted for proteasomal degradation when TORC1 is inhibited. Consistent with this possibility, Zhao *et al.*[4] found evidence for the selective proteasomal breakdown of growth-related proteins. Finally, given the integral link between ubiquitination and the proteasome, it is probable that both systems are concomitantly regulated under stress. The identification of enzymes called ubiquitin ligases and deubiquitinases, which are necessary to target substrates specifically to the proteasome, may provide a way to address this question.

From all these studies[2–4], it is clear that the TORC1/mTORC1 pathway is a central regulator of proteasome homeostasis. It will be necessary to resolve the differences in current models of how this pathway affects the proteasome, especially given that modulation of the proteasome might be a therapeutic approach for diseases such as cancer and neurodegeneration. ■

Lynne Chantranupong *and* **David M. Sabatini** *are at the Whitehead Institute, Cambridge, Massachusetts 02142, USA.*
e-mails: chantran@mit.edu; sabatini@wi.mit.edu

1. Laplante, M. & Sabatini, D. M. *Cell* **149,** 274–293, (2012).
2. Rousseau, A. & Bertolotti, A. *Nature* **536,** 184–189 (2016).
3. Zhang, Y. *et al. Nature* **513,** 440–443 (2014).
4. Zhao, J., Zhai, B., Gygi, S. P. & Goldberg, A. L. *Proc. Natl Acad. Sci. USA* **112,** 15790–15797 (2015).
5. Zhao, J., Garcia, G. A. & Goldberg, A. L. *Nature* **529,** E1–E2 (2016).
6. Finley, D. *Annu. Rev. Biochem.* **78,** 477–513 (2009).
7. Suraweera, A., Münch, C., Hanssum, A. & Bertolotti, A. *Mol. Cell* **48,** 242–253 (2012).
8. Hanssum, A. *et al. Mol. Cell* **55,** 566–577 (2014).
9. Efeyan, A. *et al. Nature* **493,** 679–683 (2012).

This article was published online on 27 July 2016.

CANCER

Fat and the fate of pancreatic tumours

In obese people with pancreatic cancer, the many interactions between fat cells and the inflammatory microenvironment surrounding the tumour lead to below-average prognosis and chemotherapy outcome.

MELEK CANAN ARKAN

The increasing prevalence of obesity will have an even greater effect on the health-care system than previously predicted, because obesity turns out to be a major risk factor for the development of cancer[1]. Obese individuals have a substantially elevated risk for a type of pancreatic cancer known as pancreatic ductal adenocarcinoma, which is the fourth most-common cause of cancer-associated death[1]. An inflammatory microenvironment is a hallmark of cancer, but little is known about how alterations in the surrounding connective tissue (stroma) contribute to tumour initiation and progression in obesity. Writing in *Cancer Discovery*, Incio *et al.*[2] report their investigation into how fat cells in the microenvironment surrounding cancer cells contribute to tumour initiation and progression in both mice and humans.

Tumour formation in the pancreas involves striking structural distortion of tissue, which is attributed to the disruption of digestive-enzyme-containing acinar cells, tissue infiltration by immune cells, a strong fibrotic response (also known as fibrosis, the formation of excess connective tissue or collagen protein around the tumour), and a higher than usual level of deposition of extracellular-matrix material. Cancer lesions in obese individuals are commonly associated with increased fat-cell (adipocyte) content compared with tumours from non-obese patients; however, the function of these fat cells in pancreatic cancer remained unclear until now.

Incio and colleagues show that, in mice, adipocytes, along with immune cells and pancreatic stellate cells, signal through the IL-1β protein and the AT1 angiotensin receptor to drive migration of immune cells called neutrophils to the tumour microenvironment. This increases the inflammatory and fibrotic response in the pancreatic-cancer microenvironment in a way that results in poor response to chemotherapy and poor prognosis.

In obese mice, the tumour microenvironment was shown to contain adipocytes that are increased in both size and number, partly as a result of tumours invading the neighbouring white adipose tissues. The researchers observed an abundant fibrotic response in tumour areas that were enriched in adipocytes or located adjacent to adipose tissue. These results suggest that fibrosis is a hallmark of adipose tissue in obese subjects with pancreatic cancer, and that the accumulation of the extracellular-matrix protein collagen, a component of the fibrotic response, in the vicinity of fat cells is a prominent characteristic of obesity. Incio and colleagues also found that adipocyte infiltration into the tumour microenvironment correlates with worse prognosis and treatment outcome in patients.

The authors hypothesized that, in people with pancreatic cancer, obesity-associated adipocyte accumulation increases fibrosis, promotes tumour progression and hinders the delivery and efficacy of chemotherapeutics. When they checked the percentage of perfused blood vessels in a given area of mouse tumour, they found that it was significantly reduced in obese animals. To determine whether impeded perfusion through blood vessels is responsible for inefficient delivery of chemotherapeutic agents, the authors measured the uptake of the chemotherapy drug 5-fluorouracil in mice. Obesity significantly decreased tumour uptake of the drug compared with uptake in non-obese control animals, thereby reducing the chemotherapy's efficacy (Fig. 1).

Chronic fibrosis is thought to have a crucial role in enhancing tumour growth and in attenuating drug delivery. However, in previous studies, inhibition of chronic fibrosis by either inhibitor compounds[3] or genetic mutations[3,4] resulted in increased immunosuppression, accelerated tumour growth and decreased survival, implying that tumour stroma may be restrictive to tumour growth.

By contrast, Incio *et al.* show that inhibition of the major pro-fibrotic pathway of AT1 signalling in mice inhibited tumour progression. The authors propose that migration into the tissue of tumour-associated neutrophils and IL-1β production are leading drivers in the regulation of tumour growth in this context, although changes in vascular perfusion due to reduced blood pressure also play a minor part. When the authors depleted neutrophils or blocked the activity of IL-1β using antibody treatment, the immunosuppressive microenvironment was reshaped and the progression of pancreatic cancer was reduced.

50 Years Ago

'We wuz robbed' — The World Cup which has recently been enacted in Britain may have been fun to watch, but there is no question that it was a thoroughly badly designed experiment … The mere fact that a Poisson distribution can describe so well the distribution of scores by individual teams goes a long way to suggest that the teams were much of a muchness in talent and their scores were independent of each other. From this point of view, the decision that the outcome of the whole competition should depend on the outcome of a single game between the two so-called finalists was as much of a farce as a great many West German supporters already know it to have been … If, for example, it were agreed that … no team should be declared the winner until its score exceeds that of its opponent by three standard deviations of Poisson distribution, it might be necessary to design the game of football so that it would be practicable for one side to score 100 goals or so … Such a change could easily be brought about, possibly by widening the goalposts or by abolishing goalkeepers.

From *Nature* 13 August 1966

100 Years Ago

The History of the Family. By Prof. W. Goodsell — In what sense is it right to speak of the history of the family? … Can it be said to have a history? … Some such questions as these arise in one's mind as one takes up Prof. Goodsell's book … even a casual reader will be struck by a want of precise references in certain of the chapters … Where is the "weight of evidence" which shows that polygamy is unpopular among savage women? The author gives several reasons why *we* condemn it, but there is surely room for doubt …

From *Nature* 10 August 1916

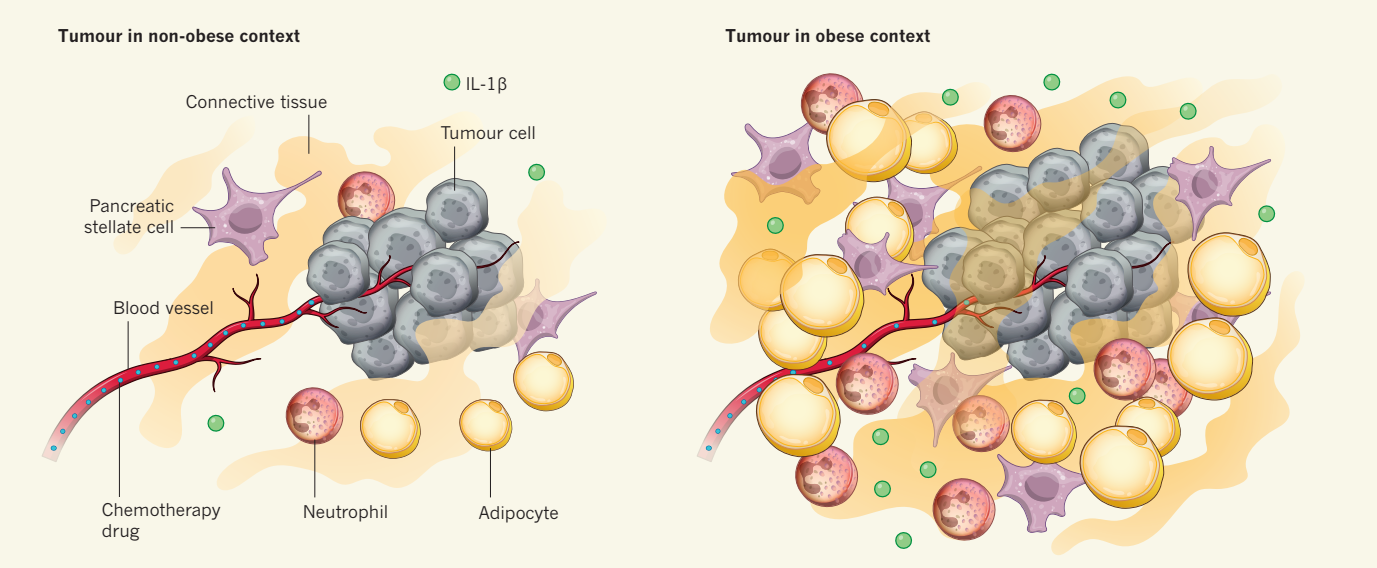


Figure 1 | Fat cells remodel the microenvironment around tumours. Tumours are perfused with blood vessels, which allow chemotherapy drugs to enter. Incio *et al.*[2] report that, in the context of obesity, access to pancreatic tumours is restricted by poor tumour blood-vessel perfusion, leading to a decreased response by tumour cells to chemotherapeutic drugs. In obesity, there is an increase in pancreatic stellate cells, immune cells such as neutrophils and IL-1β signalling molecules, as well as larger fat cells (adipocytes).The denser cellular microenvironment seen in obesity puts extra mechanical tension on the tissue and may restrict blood-vessel perfusion. This mechanical tension arises because of the signalling crosstalk between adipocytes, neutrophils, pancreatic stellate cells and other components of the tissue microenvironment. This crosstalk leads to an increase in inflammatory cells such as neutrophils and excess fibrous connective tissue in the vicinity of the tumour.

When experimentally targeting AT1 signalling in mice, other processes downstream of AT1 signalling — such as the epithelial-to-mesenchymal cell transition or adipocyte differentiation — might also be affected, and it is possible that these processes are responsible for the decrease in obesity-associated tumour progression.

Even though the authors correlated fibrotic response with tumour size, it is difficult to judge whether the fibrotic response or tumour growth comes first, because decreased tumour progression will eventually result in decreased immune-cell infiltration into the tumour microenvironment and decreased fibrosis. Either way, the results of this study and others[3,4] reinforce the need for further evaluation of the functional contribution of fibrosis in the initiation and progression of pancreatic cancer, especially in obesity.

Cellular alterations induced by mechanical forces are becoming more widely recognized as having a role in various diseases[5]. Homeostasis in the balance between internal and external forces on cells (the state of physical tension known as tensional homeostasis) can regulate apoptotic cell death, cell proliferation, adhesion and migration, and its deregulation could result in increased susceptibility to cancer. In addition, physical cues from the pressure exerted by solid-tissue components of the tumour microenvironment can compress blood vessels, causing poor tumour perfusion[6].

Incio and colleagues showed that treatment with the AT1 blocker losartan can reduce mechanical stress on cells and decrease tumour growth in mice with pancreatic cancer. More research is needed to investigate the type of transcriptional switch that tensional homeostasis induces in the dense cellular milieu of the tumour microenvironment in obesity. Inhibition of the mechanical forces acting on cells in pancreatic cancer may provide further clues for future clinical treatment. Normalizing the tumour extracellular matrix by reducing matrix stiffness may be more effective and safer than trying to delete stromal components directly.

The relationship between fat cells and stem-cell regulation is another key question. Mature white adipocytes — fat stores that control energy metabolism — respond to nutritional and hormonal cues through the secretion of signalling proteins. Studies point to white adipose fat having functions in tissue regeneration and stem-cell regulation, placing fat cells at the centre of multiple aspects of cancer progression[7]. Mesenchymal stem cells or stromal stem cells contribute substantially to adipocyte generation. Mechanical stress is also a trigger for the expansion of some stem-cell populations[8].

It would be interesting to determine the origin of adipocytes in the pancreas, their fate and phenotype (whether the cells form white or brown fat, as well as the type of component they secrete). Other potential topics for investigation include studying the contribution of adipocyte-derived signals that recruit and possibly polarize immune-cell differentiation; the function of adipocyte invasion during tumour formation; and the role of malfunctioning energy metabolism in obesity.

Incio and colleagues' work provides a plausible cellular and molecular explanation for increased adipocyte interaction with the cells of the pro-inflammatory and pro-fibrotic tumour microenvironment that accelerates disease progression and hampers therapy. Is systemically targeting tumour-associated neutrophils, pancreatic stellate cells or adipocytes feasible without causing collateral damage to host functions? Such damage could be a major challenge to successful direct translation of the current findings to the clinic. Only time will tell whether targeting IL-1β or neutrophils could offer opportunities for successful therapeutic intervention. Of course, the best preventive approach in the meantime is to eat a healthy diet and to exercise. ■

Melek Canan Arkan *is at the Institute of Biochemistry II, Goethe University, Frankfurt 60590, and the Institute for Tumor Biology and Experimental Therapy, Georg-Speyer Haus, Frankfurt 60596, Germany.*
e-mail: arkan@med.uni-frankfurt.de

1. Giovannucci, E. & Michaud, D. *Gastroenterology* **132,** 2208–2225 (2007).
2. Incio, J. *et al. Cancer Discov.* http://dx.doi.org/10.1158/2159-8290.CD-15-1177 (2016).
3. Rhim, A. D. *et al. Cancer Cell* **25,** 735–747 (2014).
4. Özdemir, B. C. *et al. Cancer Cell* **25,** 719–734 (2014).
5. DuFort, C. C., Paszek, M. J. & Weaver, V. M. *Nature Rev. Mol. Cell Biol.* **12,** 308–319 (2011).
6. Provenzano, P. P. & Hingorani, S. R. *Br. J. Cancer* **108,** 1–8 (2013).
7. Shook, B. *et al. Annu. Rev. Cell Dev. Biol.* http://dx.doi.org/10.1146/annurev-cellbio-111315-125426 (2016).
8. Guilak, F. *et al. Cell Stem Cell* **5,** 17–26 (2009).

This article was published online on 3 August 2016.

ARTICLE

doi:10.1038/nature19078

Synergistic, ultrafast mass storage and removal in artificial mixed conductors

Chia-Chin Chen[1], Lijun Fu[1] & Joachim Maier[1]

Mixed conductors—single phases that conduct electronically and ionically—enable stoichiometric variations in a material and, therefore, mass storage and redistribution, for example, in battery electrodes. We have considered how such properties may be achieved synergistically in solid two-phase systems, forming artificial mixed conductors. Previously investigated composites suffered from poor kinetics and did not allow for a clear determination of such stoichiometric variations. Here we show, using electrochemical and chemical methods, that a melt-processed composite of the 'super-ionic' conductor $RbAg_4I_5$ and the electronic conductor graphite exhibits both a remarkable silver excess and a silver deficiency, similar to those found in single-phase mixed conductors, even though such behaviour is not possible in the individual phases. Furthermore, the kinetics of silver uptake and release is very fast. Evaluating the upper limit set by interfacial ambipolar diffusion reveals chemical diffusion coefficients that are even higher than those achieved for sodium chloride in bulk liquid water. These results could potentially stimulate systematic research into powerful, even mesoscopic, artificial mixed conductors.

Mixed conductors form an important class of functional solids. They are relevant as prototype solids, of which purely electronic conductors and purely ionic conductors are special subcases, and allow for rapidly transducing chemical signals and permeating chemical components[1–4]. Therefore, they are vital in a technological context for use as electrodes, permeation membranes, sensors and catalysts. In the field of solid-state chemistry, mixed conductors are important because they enable rapid solid-state reactions; in solid-state physics, they have become popular in conjunction with the advent of high-temperature superconductivity.

Mixed conductors are characterized by a key thermodynamic parameter and a key kinetic parameter. The former is the chemical capacitance (C^{δ}, indicating a capacitance enabled by non-stoichiometry), which measures the ability of the conductor to take up or release chemical components such as oxygen, hydrogen, lithium and silver (Supplementary Information section I). The latter is the chemical diffusion coefficient (D^{δ}), which measures, for a given geometry and driving force, the rate of such processes[5], and represents the most important parameter in the field of chemical kinetics of solids. Moreover, it is typically the decisive quantity in battery research, when referring to practical energy densities. In addition to the chemical capacitance, D^{δ} also depends on the chemical resistance (R^{δ}), which itself is composed of contributions from ionic and electronic conductivities and, as such, determines the permeation rate of a component in steady-state[6].

A characteristic feature of mixed conductors is their ability to exhibit varied stoichiometry, as observed in, for example, intercalation electrodes for Li- or Na-based batteries[7], permeation membranes[8], chemical sensors[9], high-temperature superconductors[10] and materials for resistive switching[11,12]. Examples of previously investigated mixed conductors are silver chalcogenides (Ag_2Y, Y = S, Se, Te; refs 13–15). In these examples, silver excess ($Ag_{2+\varepsilon}Y$, with $\varepsilon > 0$) and silver deficiency ($\varepsilon < 0$) is realizable. Mechanistically, the former is achieved by incorporating silver ions on interstitial sites that are compensated by excess electrons and the latter by forming silver ion vacancies that are compensated by electron holes; the kinetics occurs via motion of ionic and electronic carriers coupled through charge conservation. Such stoichiometry variations occur rapidly only at elevated temperatures, especially for the highly conducting high-temperature phases[13–21]. In pure ionic conductors—such as purely Ag^+-conducting silver halides or the 'super-ionic' conductor $RbAg_4I_5$—and pure electronic conductors, perceptible stoichiometry changes are not possible, owing to the lack of one necessary carrier. This is not just a kinetic issue—in $RbAg_4I_5$, for example, the high electronic energy forbids the accommodation of a large number of excess electrons or electron holes[22].

Owing to the lack of mixed conductors with high electronic and ionic conductivities at room temperature, it is tempting to create 'artificial' or heterogeneous mixed conductors by fabricating composites consisting of compatible phases that are highly ionically and electronically conducting. Fast steady-state transport (of a component such as silver or oxygen) can be realized in a straightforward manner in bi-continuous composites of ion and electron conductors (for example, a ceramic and a metallic material) because the transport pathways can be spatially separated (dual-phase transport)[23,24]; however, enabling extra storage of matter in composite materials is delicate, because it relies on contact phenomena. These phenomena have been studied in the context of introducing a lithium, or even a hydrogen, excess in Li_2O:Ru or LiF:Ni composites[25,26] via 'job-sharing'—whereby Li^+ or H^+ is stored on the salt side of the contact and e^- or H^- on the metal side. However, these systems suffer from slow kinetics and the poorly defined contact chemistry.

Compositional variations due to excess charge have been discussed[27] on a general thermodynamic level; they are involved in various grain-boundary and surface phenomena in ceramics[28–33] or, more generally, when heterogeneous doping is considered as a function of stoichiometry[27,34]. Compositional changes at heterophase contacts are implicitly addressed in the field of supercapacitors. Here we put forward the concept of considering a composite of an ion conductor and an electron conductor (or more generally of two mixed conductors) as a heterogeneous mixed conductor that can take up or release components reversibly. Our approach is far more general than the supercapacitive view because (i) it allows for a thermodynamic and kinetic unification of phenomena such as stoichiometric variations in heterogeneous systems, (ii) it allows us to formulate the transition from a supercapacitive situation to a homogeneous bulk situation via

[1]Max Planck Institute for Solid State Research, Heisenbergstraße 1, 70569 Stuttgart, Germany.

mesoscopic intermediates[7], (iii) it allows us to couple the formalism of defect chemistry to the problem of boundary composition, and (iv) it highlights the possibility of fast lateral interfacial chemical diffusion.

Here we study the $RbAg_4I_5$:graphite composite as a prototype of an artificial mixed conductor. $RbAg_4I_5$ is a super-ionic conductor, with extremely high silver-ion conductivity and negligible electronic or anion conductivity[22]. Graphite is a purely electronic conductor with the ability to exhibit n- and p-type transport[35,36]. The interfacial polarization of silver halides and alkali silver halides in contact with graphite has been studied previously[37–40], but the storage behaviour has not yet been explored. The artificial mixed conductor we study exhibits two anomalies. First, unlike the pure phases that are stoichiometrically inert, the mixed conductor allows for not only a substantial silver excess, but—characteristically—also a silver deficiency; therefore, it effectively shows a veritable homogeneity range, similarly to a mixed conducting bulk phase, which enables a generalized treatment of stoichiometric changes in interfacially controlled materials. Second, this stoichiometric variation propagates very rapidly along the interface. Even though the transport kinetics for bi-continuous phases takes advantage of bulk migration pathways, it is the interfacial transport path that is conceptually most intriguing, for the following reasons: (i) it sets an upper limit to the relaxation time; (ii) it is characterized by an ambipolar diffusion process, unlike the bulk path, and so the chemical diffusion coefficients can be compared with the corresponding values in pure phases; and (iii) it determines the kinetics of the non-compacted composites that we study. If such composites are used as components of electrochemical cells, then very high rate performances are achieved, supported by the fact that the super-ionic conductor can naturally be used as a solid electrolyte and that usual passivation problems do not arise.

Equilibrium stoichiometric variation

To add or remove silver we use the coulometric cell $Ag/RbAg_4I_5/RbAg_4I_5$:graphite/Pt. By passing a current through the cell for a given length of time, silver titration (by charge Q) of the composite is achieved (Fig. 1). The limit of silver addition is reached when silver is deposited on the Pt side and the open-circuit cell voltage (E) is zero; the limit of any silver removal is indicated by the oxidation of the iodine ion (that is, liberation of I_2), which starts at a voltage of approximately 470 mV, consistent with observations[37] and with the standard decomposition voltage[41] of approximately 670 mV. For pure $RbAg_4I_5$, the change between these two limiting values (0 and 470 mV) is abrupt, which indicates negligible stoichiometric variation of Ag, in correspondence with the extremely low electronic-carrier concentration in the material. In the case of carbon, the $E(Q)$ curve exhibits an abrupt change from high, not-well-defined values to zero, which also indicates zero silver storage. The result for the composite is markedly different. Figure 1a shows the expected transition from the I_2 liberation voltage to zero voltage, but with substantial stoichiometric variation in between. Silver excess and deficiency are realized, as in usual mixed conductors (for example, Ag_2S, Ag_2Se and Ag_2Te)[13,17,18]. In contrast to these mixed conductors, the stoichiometric zero point of the charge–voltage curve of our composite is well defined because it refers to the starting point after putting two very stoichiometric (on the scale of interest) phases into contact (Supplementary Information section I and Supplementary Fig. 3).

The silver excess is realized by silver interstitials ($Ag_i^{\bullet}$) on the ionic conductor side and excess electrons (e') on the electronic conductor side, whereas the deficiency is realized by silver vacancies (V'_{Ag}) on one side and electron holes ($h^{\bullet}$) on the other side (Fig. 1b). We briefly discuss the theoretical background in Supplementary Information section I. A quantitative and comprehensive thermodynamic treatment of the stoichiometric variation of these job-sharing composites is beyond the scope of this Article. It can be shown that the effect of added or removed charge on the voltage can generally be split into a contribution from the diffuse double layer and a contribution from the potential jump over the interface[25], with the first being negligible, owing to the high bulk ionic- and electronic-charge-carrier contributions, leading to the rather linear E–Q characteristic.

Stoichiometric effects increase with the contact area between the two materials, and are zero for the pure constituents (Fig. 1c). Dispersions of graphite in $RbAg_4I_5$ and of $RbAg_4I_5$ in graphite lead to an increase

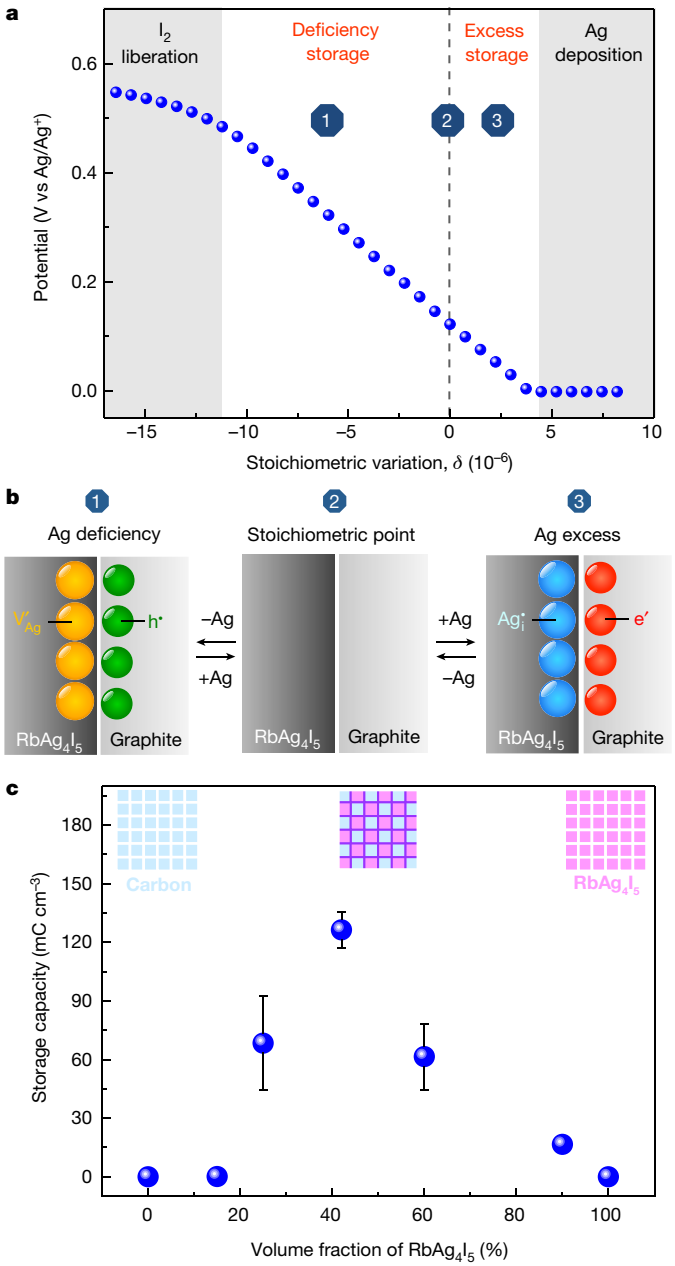


Figure 1 | Silver excess and deficiency through interfacial storage. **a**, Variation of silver content in a $RbAg_4I_5$:graphite (90%:10%) composite, achieved by coulometric titration. The potential range is bound by decomposition ('I_2 liberation', left) and Ag deposition (right) without indication of underpotential deposition. **b**, Schematic of interfacial Ag deficiency and excess. On the ionic-conductor ($RbAg_4I_5$) side, the deficiency is realized by Ag vacancies (V'_{Ag}) and the excess by interstitials ($Ag_i^{\bullet}$); on the electronic-conductor (graphite) side, the deficiency is realized by electron holes ($h^{\bullet}$) and the excess by excess electrons (e'). **c**, Storage capacity as a function of volume fraction, measured by coulometric titration between 10 mV and 400 mV. The composite (centre inset) shows distinct storage capacity, unlike the pure phases (carbon, left inset; $RbAg_4I_5$, right inset). The position of the maximum varies if different normalization is used (for example, per mole number), but the general trend remains that the capacity increases with the contact area between the different phases. Error bars show range over three samples.

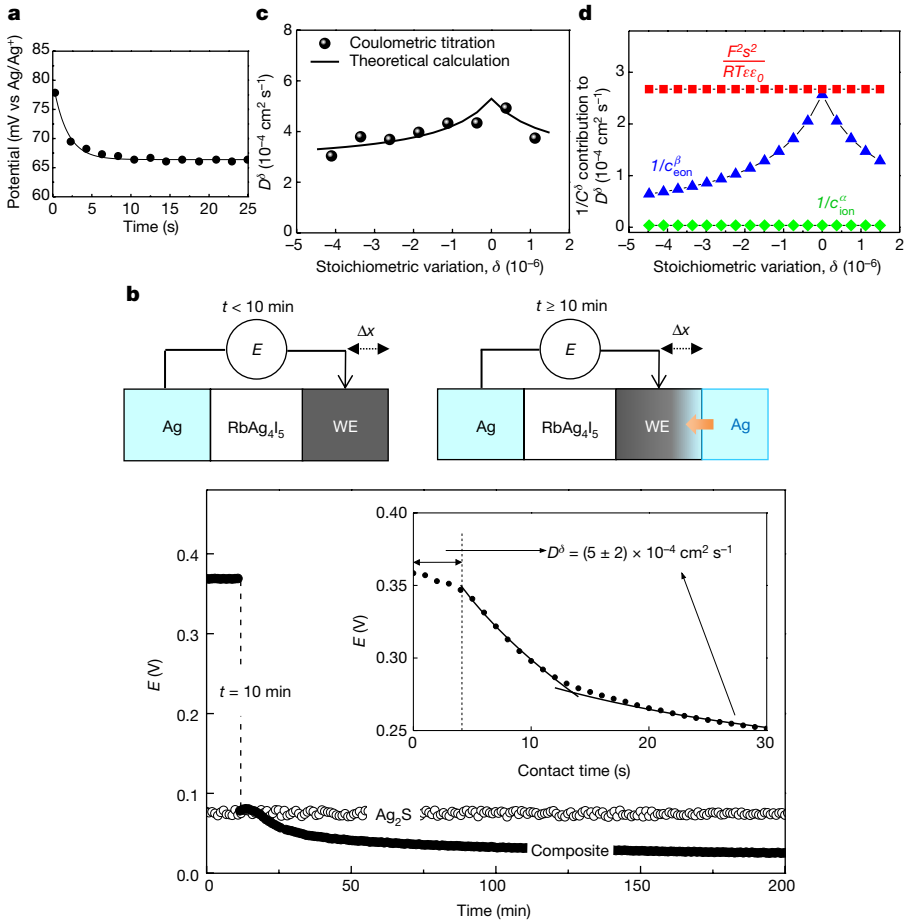


Figure 2 | Rapid silver dissolution in the job-sharing composite. a, The relaxation process of the coulometric titration shows anomalously rapid silver storage/removal in the $RbAg_4I_5$:graphite composite. **b**, Ag permeation experiment. At $t = 10$ min, the right-hand working electrode (WE) is suddenly brought into contact with silver. (Blue parts of the cells refer to high Ag potential.) For the 'job-sharing' composites, the open-circuit cell voltage (E, measuring the silver chemical potential at Δx) approaches zero very quickly. Unlike $RbAg_4I_5$:graphite (filled circles), in the case of Ag_2S (open circles) E has not changed even after 2 h. The inset refers to an experiment that was specially designed to follow the transient and enable evaluation of the effective diffusion coefficient. (The dashed line in the inset marks the time lag. The evaluation of this time lag as well as the evaluation of the long-time behaviour yield D^δ with a tolerable deviation indicated in the brackets.) **c**, Comparison of the chemical diffusion coefficient D^δ determined from titration data (circles) and theoretical calculations (solid line). The prediction by the job-sharing model is in excellent agreement with the data. **d**, Contributions of various capacitive terms to D^δ ($1/C^\delta$ is the reciprocal chemical capacitance). The characteristic term resulting from the electrostatic field effect (red squares) is dominant and increases the chemical diffusion in the composite to unprecedentedly high values. For definition of the quantities, see Table 1.

in the capacity, with the maximum achieved at 40 vol% $RbAg_4I_5$. The correlation with the interfacial area is also highlighted by our results on $RbAg_4I_5$ composites with multi-wall carbon nanotubes (Supplementary Fig. 10). Note that the δ-value in Fig. 1 averages over the whole coarse-grained sample. The local δ-value is very high and corresponds to a silver excess of about 15% in the monolayer where the phases contact one another (on the basis of Brunauer–Emmett–Teller measurements; Supplementary Information section I).

Kinetics of silver storage and removal

Equally exciting as the finding of a compositional variation ranging from deficiency to excess is the kinetics. A typical transient of the coulombic titration is given in Fig. 2a. By considering the composite as an effective mixed conductor, an effective diffusion coefficient is derived that is extremely high and, for the non-compacted composite, is $4 \times 10^{-4}\,cm^2\,s^{-1}$ (Supplementary Information section II). This value agrees well with that predicted from the analysis of job-sharing diffusion along the interface. The minor importance of bulk migration (followed by double-layer charging) is understandable from the partly mesoscopic microstructure and the wetting behaviour, as detailed in Supplementary Information section III. Such rapid changes in stoichiometry at room temperature are unprecedented, and the chemical diffusion coefficients for the interfacial transport exceed bulk values for known systems by orders of magnitudes. Owing to the slow room-temperature kinetics, there are virtually no room-temperature measurements of chemical diffusion coefficients, except for lithium diffusion in batteries. The values reported there[42] are more than five orders of magnitude lower than those for our super-ionic conductor/graphite composite. Also, the

Table 1 | Chemical resistance cells and chemical capacitance for chemical diffusion cells

	$1/R^\delta$	$1/C^\delta$
Classic bulk chemical diffusion $\alpha = Ag_2Te$, $LiFePO_4$, $SrTiO_3$ and so on	$\frac{\sigma^\alpha_{ion}\sigma^\alpha_{eon}}{\sigma^\alpha_{ion}+\sigma^\alpha_{eon}}$	$\frac{1}{c^\alpha_{ion}}+\frac{1}{c^\alpha_{eon}}$
Job-sharing chemical diffusion $\alpha = RbAg_4I_5$, β = graphite	$\frac{\sigma^\alpha_{ion}\sigma^\beta_{eon}}{\sigma^\alpha_{ion}+\sigma^\beta_{eon}}$	$\frac{1}{c^\alpha_{ion}}+\frac{1}{c^\beta_{eon}}+\frac{F^2s^2}{RT\varepsilon\varepsilon_0}$

The product of chemical resistance R^δ and chemical capacitance C^δ yields the relaxation time $\tau^\delta \propto 1/D^\delta$ (proportionality constant given by L^2). The expression given for $1/C^\delta$ is valid only for the dilute approximation and in the absence of trapping. It is necessary to introduce activity coefficients as well-known thermodynamic corrections in the case of high charge-carrier concentration (Supplementary Information section IV), in particular for $RbAg_4I_5$. Because of the dominance of the third term in the expression for $1/C^\delta$ for the composite, the qualitative and quantitative conclusions are not affected by such correction factors. $\sigma_{ion(eon)}$, ionic (electronic) conductivity; $c_{ion(eon)}$, ionic (electronic) concentration; s, separation distance of the atoms; F, Faraday's constant; R, gas constant; T, temperature; $\varepsilon\varepsilon_0$, absolute dielectric constant.

Table 2 | Comparison of chemical diffusion coefficients D^{δ} at 25 °C

	$RbAg_4I_5$/graphite composites			$RbAg_4I_5$	NaCl in water
Source	Coulometric titration	Ag permeation	Calculation	Ref. 49	Ref. 48
D^{δ} (10^{-5} cm^2 s^{-1})	30–50	50 ± 20	30–60	0.0017	1.6

The indicated range of the experimental results for the coulometric titration and of the calculations refers to the entire stoichiometry range (Fig. 2c). The error bars concerning the experimental Ag-permeation data represent the range of values from two evaluation procedures (inset, Fig. 2b).

values for high-temperature diffusors when extrapolated to room temperature are distinctly smaller than our value (Fig. 3b).

The very high diffusion coefficients for our composite are due to the very low chemical resistance compared to the pure constituents (ions take the ionic pathways in $RbAg_4I_5$ and electrons the electronic pathways in graphite) and, more subtly, to the low (differential) chemical capacitance. The kinetics is faster than for a hypothetical bulk mixed conductor with same charge-carrier concentrations and mobilities. The reason for the fast kinetics is the third term in the equation for $1/C^{\delta}$ in the second row of Table 1, which refers to the electrostatic energy that needs to be overcome if the concentration is changed at the heterojunction. As shown in Fig. 2c, d, this term dominates the other two terms in this equation by an order of magnitude. This dominance is even more evident if we include 'non-ideality' terms[43] (Supplementary Information sections III, IV). The $1/c_{eon}$ term is as important as the third term only at the point of zero charge, leading to the small maximum in D^{δ} seen in Fig. 2c, d. The calculation of D^{δ} using known material parameters yields 5×10^{-4} cm^2 s^{-1}, in excellent agreement with the experimental value of the chemical diffusion coefficient (Table 2).

This value of D^{δ} also follows formally from a supercapacitive (de Levie-type) model[44] in which the boundaries are charged from the bulk side by shrinking the bulk phases to single layers. The electrochemical capacitance is then expressed in terms of double-layer charging with the chemical part stemming from the configurational entropy of the charge on the capacitor plates (Supplementary Information section III). In the case of a bi-continuous composite with major contributions from bulk migration, the problem is more complicated because of the greatly increased dimensionality and so needs to be addressed using finite element modelling. (The relaxation time for ideal bulk contributions, which cannot be interpreted in terms of a microscopic diffusion coefficient, is addressed below and shown to be of the order of tens of milliseconds for the grain sizes under consideration.)

The schematic in Fig. 3a illustrates the effects of chemical resistance and capacitance that contribute to the anomalously high chemical diffusion of the job-sharing composites along with the bulk-assisted supercapacitive mode (grey arrows). Table 2 highlights the fact that the observed value of D^{δ} and that derived from the ambipolar interfacial transport model are unprecedentedly high, exceeding the values for other solid room-temperature systems, and even exceeding the diffusion coefficients of NaCl in liquid water, by orders of magnitude (Fig. 3b).

Further corroboration of fast transport comes from another experiment. In this experiment, we track silver dissolution (that is, the propagation of the chemical potential of silver) using electrochemical detection, as sketched in Fig. 2b. As the initial condition, a jump in the chemical potential is created at the outer side of the composite (working electrode) by bringing it into contact with silver[45].

Figure 2b refers to a thermodynamically well-defined set-up (see Methods) and shows that the silver signal covered the distance Δx and thus reached the sensing point in less than a minute, leading to zero potential difference. The inset refers to a refined experiment that allows us to resolve the relaxation time. According to a detailed evaluation (Supplementary Information section II), this relaxation time agrees well with the diffusion coefficient given above. Such a quick response is not observed for mixed conductors such as Ag_2S (Fig. 2b) and Ag_2Se (Supplementary Fig. II(B)), as the comparison shows.

To demonstrate the usefulness of super-ionic conductor/graphite composites in electrochemical devices[37–39,46], we briefly discuss its

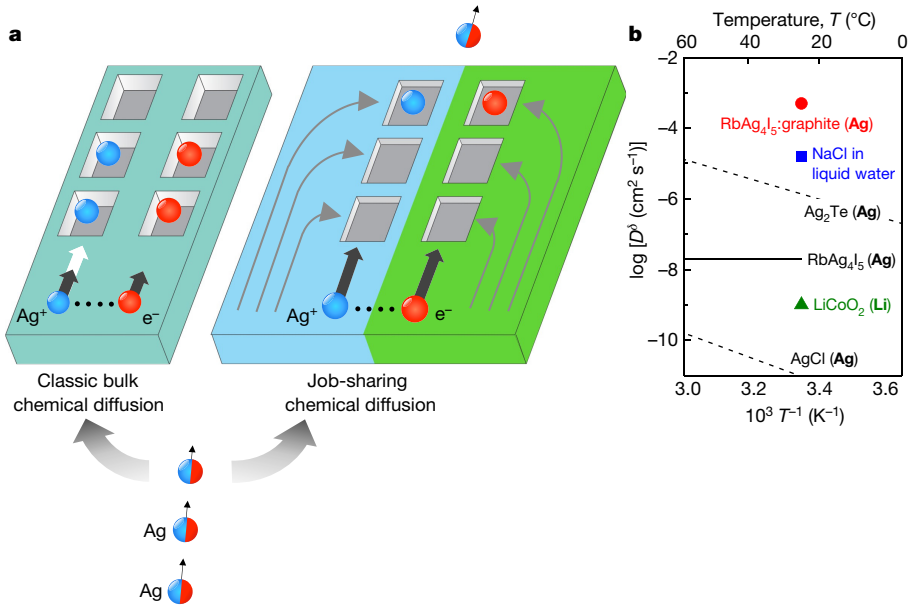


Figure 3 | Rapid job-sharing diffusion. a, Schematic of the different chemical diffusion mechanisms (bulk and job-sharing). In both cases, coupled motion occurs during which the slower component slows down the effective conductivity. The composite phases can be selected such that both contributions are high. Even for the same conductivities and carrier concentrations, the chemical diffusion is faster in the composite because the compositional change is less important, owing to the internal electrostatic energy, as indicated by the lower depths of the grey boxes. The grey arrows indicate supercapacitive charging contributions that benefit from bulk migration and are dominant for bi-continuous bulk percolation. **b**, Comparison of chemical diffusion coefficients for different materials at room temperature (dashed lines are extrapolated from ref. 14 (Ag_2Te) or ref. 50 (AgCl)). Values for $RbAg_4I_5$ are from ref. 49. The value for our composite (red circle) exceeds the reported values for other solids (including $LiCoO_2$ (ref. 42), green triangle) and even for NaCl in liquid water[48] (blue square) by at least one order of magnitude. (Bold elements in brackets indicate the transported elements.)

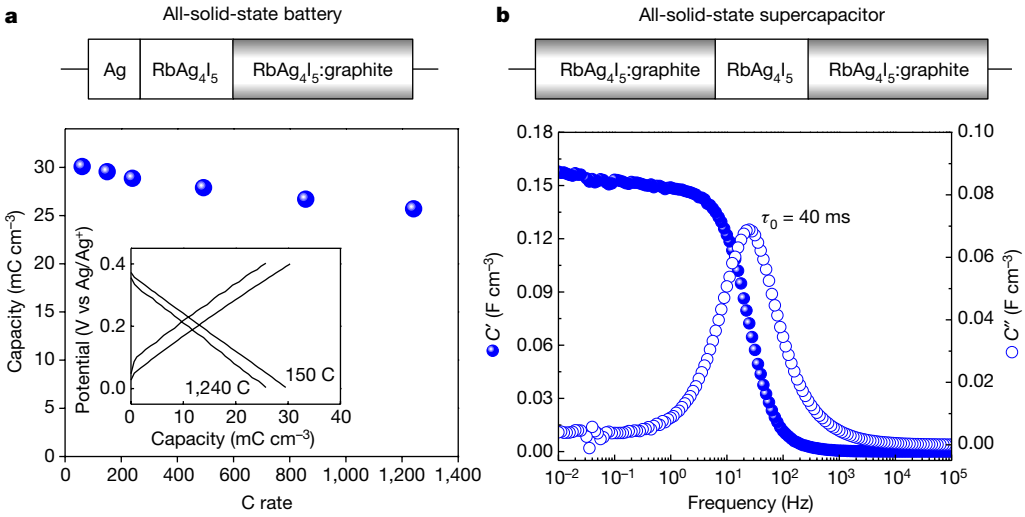


Figure 4 | Ultrafast performance of electrochemical devices using $RbAg_4I_5$:graphite. a, The capacity of an all-solid-state battery at different C rates. As per common practice, a C rate of 1 denotes the rate at which a full charge or discharge takes 1 h; for example, it takes 3 s at 1,200 C. Inset, voltage-capacity curves. **b**, All-solid-state supercapacitor. Here, we realize silver excess (interstitials on the $RbAg_4I_5$ side, electrons on the graphite side) on one side of the cell and silver deficiency (vacancies on the $RbAg_4I_5$ side, holes graphite side) on the other. The frequency dependence of the real (C'; left axis) and imaginary (C''; right axis) parts of the capacitance shows an extremely short relaxation time of $\tau_0 = 40$ ms, indicating ultrafast charge/discharge rates. The short relaxation time also suggests that local supercapacitive action is assisted by interfacial neutral mass transport (redistribution of the polarized state).

performance in two applications. For these experiments, we prepared samples with a maximum of bulk percolation by applying substantial pressure to the samples during the preparation. In the first experiment (Fig. 4a), the micrometre-sized composite is used as an electrode in a battery cell and can be reversibly charged at extremely high rates. In this case, we are able to use the entire effective stoichiometric width of the composite. In the second experiment (Fig. 4b), the composite is used within a supercapacitor cell, which has a relaxation time characterized from impedance spectroscopy of 40 ms. We are not aware of a quicker relaxation process for an all-solid-state supercapacitor device. Even devices based on liquid systems exhibiting such values are considered to be extremely fast[46]. Note that here we refer to silver excess ($Ag_i^{\bullet}/e'$) on one side and to silver deficiency ($V'_{Ag}/h^{\bullet}$) on the other. The observed relaxation time is close to the ideal supercapacitive response time (Supplementary Information section III), indicating essentially bulk percolation; however, owing to the non-ideal microstructure, favourable bridging effects, according to the described interfacial diffusion over short distances, can be assumed to occur. It can be conjectured that, in many supercapacitors (particularly those with a partial covalency or partial Faradaic process), the lack of such a possibility could be a reason for decreased practical rate performance. Details of such studies are beyond the scope of this Article.

If both bulk and interfacial transport are macroscopically relevant, then we can conclude that it is the upper limit of the measured relaxation time that is determined by the interfacial diffusion and thus by the component redistribution kinetics along the percolating interfaces of the composites. In the context of supercapacitive action, beyond the quick build-up of boundary polarization via bulk transport (charging of capacitor plates from outside)[37–39], the rapid propagation of this electrochemical polarization state along the boundaries is a key issue (transport along the capacitor plates). These boundary effects will become more dominant for composites of weak conductors where migration processes in the bulk are less important and, in particular, in systems in which the species of interest is not conductive at all in the bulk (such as dissociative hydrogen incorporation in Li_2O-Ru composites[26]). A preference of transport along the interface as compared to lateral charging is also realized in strongly anisotropic materials (such as graphite) and in thin film systems of nanometre-scale dimensions as partly realized in our non-compacted composites (Supplementary Information section III). Such effects will be highly relevant for targeted future artificial conductors with extreme storage capacities (monolayer heterostructures of fast ionic and electronic conductors), for chemical transport along grain boundaries in ceramics and for surface processes in catalysis (such as spill-over).

Conclusion

Here we show that powerful mixed conductors can be constructed by forming appropriate composites of super-ionic conductors and pure electronic conductors, displaying a proper homogeneity range with excess and deficiency. This homogeneity range enables the design of new mixed conductors, for example, those without redox-active elements, and generalization of the stoichiometry concept for heterogeneous systems. The local storage capacity can be very high, promising pronounced storage capacities in adequately nanostructured devices. Moreover, a unified treatment of mass-transport kinetics in heterogeneous media is enabled by our approach. Storage in such composites can be extremely fast, owing to bulk migration and the astonishingly fast chemical diffusion along the boundaries. This interfacial diffusion path will be decisive for future artificial mixed conductors based on mesoscopic phase constituents.

We believe that this study will stimulate systematic research on artificial, fast, mixed conductors with substantial potential for supercapacitor systems, electrodes, permeation membranes or catalysts, by constructing relevant nanometre-scale heterostructures in analogy to artificial ion-conductor systems[47]. Besides the practical aspects, we emphasize the conceptual broadening achieved by considering thermodynamics and kinetics of heterogeneous storage.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 October 2015; accepted 24 June 2016.

1. Kudo, T. & Fueki, K. *Solid State Ionics* Ch. 5, 47–63 (VCH, 1990).
2. Tuller, H. L. Semiconduction and mixed ionic–electronic conduction in nonstoichiometric oxides: impact and control. *Solid State Ion.* **94,** 63–74 (1997).
3. Riess, I. Mixed ionic–electronic conductors—material properties and applications. *Solid State Ion.* **157,** 1–17 (2003).
4. Maier, J. Nanoionics: ion transport and electrochemical storage in confined systems. *Nat. Mater.* **4,** 805–815 (2005).
5. Wagner, C. Equations for transport in solid oxides and sulfides of transition metals. *Prog. Solid State Chem.* **10,** 3–16 (1975).

6. Jamnik, J. & Maier, J. Generalised equivalent circuits for mass and charge transport: chemical capacitance and its implications. *Phys. Chem. Chem. Phys.* **3,** 1668–1678 (2001).
7. Maier, J. Thermodynamics of electrochemical lithium storage. *Angew. Chem. Int. Ed.* **52,** 4998–5026 (2013).
8. Bouwmeester, H. J. M. & Burggraaf, A. J. in *The CRC Handbook of Solid State Electrochemistry* (eds Gellings, P. J. & Bouwmeester, H. J. M.) 481–554 (CRC Press, 1997).
9. Kim, I.-D., Rothschild, A. & Tuller, H. L. Advances and new directions in gas-sensing devices. *Acta Mater.* **61,** 974–1000 (2013).
10. Maier, J. & Pfundtner, G. Defect chemistry of the high-T_c superconductors. *Adv. Mater.* **3,** 292–297 (1991).
11. Waser, R. & Aono, M. Nanoionics-based resistive switching memories. *Nat. Mater.* **6,** 833–840 (2007).
12. Nilges, T. *et al.* Reversible switching between p- and n-type conduction in the semiconductor $Ag_{10}Te_4Br_3$. *Nat. Mater.* **8,** 101–108 (2009).
13. Schmalzried, H. Ag_2S—the physical chemistry of an inorganic material. *Prog. Solid State Chem.* **13,** 119–157 (1980).
14. Bürgermeister, A. & Sitte, W. Chemical diffusion in β-Ag_2Te. *Solid State Ion.* **141–142,** 331–334 (2001).
15. Beck, G. & Janek, J. Coulometric titration at low temperatures—nonstoichiometric silver selenide. *Solid State Ion*. **170,** 129–133 (2004).
16. Becker, K. D., Schmalzried, H. & von Wurmb, V. The chemical diffusion coefficient in (low temperature) α-Ag_2S determined by an electrochemical relaxation method. *Solid State Ion*. **11,** 213–219 (1983).
17. Sitte, W. Chemical diffusion in mixed conductors: α'-Ag_2Te and β-Ag_2Se. *Solid State Ion*. **94,** 85–90 (1997).
18. Oehsen, U. V. & Schmalzried, H. Thermodynamic investigations of Ag_2Se. *Ber. Bunsenges. Phys. Chem* **85,** 7–14 (1981).
19. Kanai, H. *et al.* Defect chemistry of $La_{2-x}Sr_xCuO_{4-\delta}$: oxygen nonstoichiometry and thermodynamic stability. *J. Solid State Chem.* **131,** 150–159 (1997).
20. Bakken, E., Norby, T. & Stølen, S. Nonstoichiometry and reductive decomposition of $CaMnO_{3-\delta}$. *Solid State Ion*. **176,** 217–223 (2005).
21. Mueller, D. N., De Souza, R. A., Yoo, H.-I. & Martin, M. Phase stability and oxygen nonstoichiometry of highly oxygen-deficient perovskite-type oxides: a case study of (Ba, Sr)(Co, Fe)$O_{3-\delta}$. *Chem. Mater.* **24,** 269–274 (2012).
22. Funke, K. Agl-type solid electrolytes. *Prog. Solid State Chem.* **11,** 345–402 (1976).
23. Sunarso, J. *et al.* Mixed ionic–electronic conducting (MIEC) ceramic-based membranes for oxygen separation. *J. Membr. Sci.* **320,** 13–41 (2008).
24. Maier, J. in *Modern Aspects of Electrochemistry* Vol. 41 (eds Vayenas, C. *et al.*) 1–138 (Springer, 2007).
25. Fu, L., Chen, C.-C., Samuelis, D. & Maier, J. Thermodynamics of lithium storage at abrupt junctions: modeling and experimental evidence. *Phys. Rev. Lett.* **112,** 208301 (2014).
26. Fu, L. *et al.* "Job-sharing" storage of hydrogen in Ru/Li_2O nanocomposites. *Nano Lett.* **15,** 4170–4175 (2015).
27. Maier, J. Kröger–Vink diagrams for boundary regions. *Solid State Ion*. **32–33,** 727–733 (1989).
28. Hagenbeck, R. & Waser, R. Influence of temperature and interface charge on the grain-boundary conductivity in acceptor-doped $SrTiO_3$ ceramics. *J. Appl. Phys.* **83,** 2083–2092 (1998).
29. Chiang, Y. M., Lavik, E. B., Kosacki, I., Tuller, H. L. & Ying, J. Y. Defect and transport properties of nanocrystalline CeO_{2-x}. *Appl. Phys. Lett.* **69,** 185–187 (1996).
30. Kim, S. & Maier, J. On the conductivity mechanism of nanocrystalline ceria. *J. Electrochem. Soc.* **149,** J73–J83 (2002).
31. Frömling, T. *et al.* Oxygen tracer diffusion in donor doped barium titanate. *J. Appl. Phys.* **110,** 043531 (2011).
32. Kim, S., Merkle, R. & Maier, J. Oxygen nonstoichiometry of nanosized ceria powder. *Surf. Sci.* **549,** 196–202 (2004).
33. Lupetin, P., Gregori, G. & Maier, J. Mesoscopic charge carriers chemistry in nanocrystalline $SrTiO_3$. *Angew. Chem. Int. Ed.* **49,** 10123–10126 (2010).
34. Petuskey, W. T. Interfacial effects on Ag:S nonstoichiometry in silver sulfide/alumina composites. *Solid State Ion*. **21,** 117–129 (1986).
35. Tatar, R. C. & Rabii, S. Electronic properties of graphite: a unified theoretical study. *Phys. Rev. B* **25,** 4126–4141 (1982).
36. Huang, X., Qi, X., Boey, F. & Zhang, H. Graphene-based composites. *Chem. Soc. Rev.* **41,** 666–686 (2012).
37. Hull, M. N. & Pilla, A. A. The transient behavior of graphite-silver iodide and platinum-silver iodide interfaces in a solid-state system. *J. Electrochem. Soc.* **118,** 72–78 (1971).
38. Oxley, J. A solid state electrochemical capacitor. *141st Meeting of The Electrochemical Society* abstr. no. 175, p. 446 (Houston, 1972).
39. Raleigh, D. O. Electrode capacitance in silver-halide solid electrolyte cells. I. Room temperature graphite and platinum electrodes. *J. Electrochem. Soc.* **121,** 632–639 (1974).
40. Armstrong, R. D. & Horrocks, B. R. The double layer structure at the metal–solid electrolyte interface. *Solid State Ion*. **94,** 181–187 (1997).
41. Scrosati, B., Germano, G. & Pistoia, G. Electrochemical properties of $RbAg_4I_5$ solid electrolyte. I. Conductivity studies. *J. Electrochem. Soc.* **118,** 86–89 (1971).
42. Barker, J., Pynenburg, R., Koksbang, R. & Saidi, M. Y. An electrochemical investigation into the lithium insertion properties of Li_xCoO_2. *Electrochim. Acta* **41,** 2481–2488 (1996).
43. Maier, J. Mass transport in the presence of internal defect reactions—concept of conservative ensembles: I, chemical diffusion in pure compounds. *J. Am. Ceram. Soc.* **76,** 1212–1217 (1993).
44. De Levie, R. in *Advances in Electrochemistry and Electrochemical Engineering* Vol. 6 (ed. Delahay, P.) 329–397 (Wiley-Interscience, 1967).
45. Heyne, L. & Beekman, N. Electronic transport in calcia-stabilized zirconia. *Proc. Brit. Ceram. Soc.* **19,** 229–263 (1971).
46. Pech, D. *et al.* Ultrahigh-power micrometre-sized supercapacitors based on onion-like carbon. *Nat. Nanotechnol.* **5,** 651–654 (2010).
47. Sata, N., Eberman, K., Eberl, K. & Maier, J. Mesoscopic fast ion conduction in nanometre-scale planar heterostructures. *Nature* **408,** 946–949 (2000).
48. Cussler, E. L. *Diffusion: Mass Transfer in Fluid Systems* (Cambridge Univ. Press, 1997).
49. Bredikhin, S., Hattori, T. & Ishigame, M. Ambipolar diffusion in $RbAg_4I_5$. *Solid State Ion*. **67,** 311–316 (1994).
50. Mizusaki, J. & Fueki, K. Electrochemical determinations of the chemical diffusion coefficient and non-stoichiometry in AgCl. *Solid State Ion*. **6,** 85–91 (1982).

Supplementary Information is available in the online version of the paper.

Acknowledgements We acknowledge technical support by A. Fuchs (N_2 isotherms), J. Liu (SEM), H. Hoier (XRD), V. Duppel (TEM) and R. Merkle (TGA). We are grateful to C. Wu for discussions and B. Lotsch for reading the manuscript.

Author Contributions The scientific conception is due to J.M. C.-C.C. designed and executed the experiments. L.F. assisted in experiments and discussions. C.-C.C. and J.M. analysed the data, are responsible for the theoretical treatment and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.M. (s.weiglein@fkf.mpg.de).

METHODS

Synthesis of $RbAg_4I_5$ and $RbAg_4I_5$/carbon composites. $RbAg_4I_5$ was synthesized by melting a near-stoichiometric mixture (83:17 mol%) of AgI (Sigma-Aldrich, 99.999%) and RbI (Sigma-Aldrich, 99.9%) at 450 °C for 2 h, followed by rapid cooling. $RbAg_4I_5$/graphite composites were first mixed with AgI/RbI and the appropriate amount of graphite (Alfa Aesar, 99.9995%) in an agate mortar. (The water content of graphite is less than 100 p.p.m., according to thermogravimetric analysis.) To improve the $RbAg_4I_5$/graphite contact, the mixture was melted at 450 °C for 10 h. For composites of $RbAg_4I_5$ and multi-wall carbon nanotubes (MWCNTs), the procedure was the same except graphite was replaced with MWCNTs (Sigma-Aldrich, 98%; the dimensions of the nanotubes are 10 nm (outer diameter), 4.5 nm (inner diameter), and 6 μm (length). All syntheses were carried out in the dark under Ar atmosphere.

Characterization method. Structure and crystallinity were characterized by X-ray diffraction (XRD) using a Philips PW3710 (40 kV/30 mA) with Cu-Kα radiation. Scanning electron microscopy (SEM) analysis was performed using a Zeiss Gemini DSM 982 scanning electron microscope. The nitrogen adsorption and desorption isotherms were measured using an Autosorb-1 system (Quanta Chrome). Transmission electron microscopy (TEM) was performed with a Phillips CM30 ST (300 kV, LaB_6 cathode). Thermogravimetry was carried out with Netzsch STA449C Jupiter TG.

Electrochemical experiments. The prepared $RbAg_4I_5$ and $RbAg_4I_5$/graphite composites served as the solid electrolyte and cathode. A mixture of silver powder (Sigma-Aldrich, 99.9%) and $RbAg_4I_5$ in the weight ratio of 4:1 was used as the anode. Powder compacts of each cell component were uniaxially pressed into 5-mm-diameter pellets at 35 MPa for 1 min. Graphite pellets are difficult to prepare by cold pressing; hence, graphite rods (Sigma-Aldrich, 99.999%) were used instead. Platinum foil and Ag foil were used as cathodic and anodic current collectors, respectively. After assembling all the components, the batteries were placed in gas-tight quartz set-ups. To reduce the contact resistance, the cell was slightly spring-loaded. Under Ar atmosphere, the electrochemical experiments were performed in the dark at 25 °C on an Arbin MSTAT system.

The coulometric titration experiment (that is, galvanostatic intermittent titration technique, GITT) consists of a series of current pulses, followed by an equilibration period. Constant currents for specific intervals of time were then applied. Depending on the volume ratio, the time intervals ranged from 30 s to 3 min for $RbAg_4I_5$/graphite composites (10^{-5} A/g) and from 3 min to 5 min for $RbAg_4I_5$/MWCNTs composites (10^{-4} A/g). After switching off the current, the open-circuit potential was measured until the system has reached equilibrium. The equilibration time typically ranged from 15 min to 60 min.

For the extremely slow galvanostatic charge/discharge on pure $RbAg_4I_5$, a very small constant current of 4×10^{-10} A/g was supplied by a Keithley 220 current source. The potential was measured by a Keithley 6514 electrometer.

Besides battery cells, all-solid-state supercapacitors were fabricated by sandwiching a $RbAg_4I_5$ pellet in between two $RbAg_4I_5$/graphite composite pellets. The preparation methods and the pellet sizes were the same as mentioned earlier. Platinum foils were used as current collectors. For the rate evaluation, the electrochemical cells (batteries and supercapacitors) were tested in Swagelok-type cells. The assembling was done in an Ar-filled glovebox (O_2 content of <0.3 p.p.m., H_2O content of <0.1 p.p.m.). The electrochemical impedance spectroscopy was conducted by a Voltalab PGZ402 impedance analyser with 5-mV signal amplitude. The frequencies ranged from 10^5 Hz to 0.01 Hz. The relaxation time of the supercapacitor was obtained by analysing the frequency dependence of the real and imaginary parts of the complex capacitance[51].

Ag permeation experiment. Ag_2S (Sigma-Aldrich, 99.9%), Ag_2Se (Sigma-Aldrich) and $RbAg_4I_5$/graphite composites were used as working electrodes. The anode, electrolyte, anodic current collector and the pressing conditions were the same as for the electrochemical experiments. The sensor probe was similar to the electronic probe used in ref. 52. The length of the working electrode was about 2–3 mm. Pt wire was wrapped around the working electrode bar. The position of the Pt wire was in the centre of the bar.

The electromotive force (E) was measured under Ar atmosphere between the Ag anode and the working electrode. E was very stable, at least for 10 h. Before starting the permeation experiment, the quartz set-up was opened and the E measurement switched off. An Ag pellet was contacted and attached to the outer side of the working electrode. The measurement set-up was then closed and flushed with Ar. The reassembling process took about 1 min. Afterwards, E was recorded for at least 6 h.

Because the time lag of the $RbAg_4I_5$/graphite composite is shorter than the time required for reassembling the set-up, a special measurement was performed to allow us to track the transient. For this purpose, silver was attached *in situ* without interrupting the circuit. Because this was done under air, the absolute emf values are less reliable.

51. Taberna, P. L., Simon, P. & Fauvarque, J. F. Electrochemical characteristics and impedance spectroscopy studies of carbon-carbon supercapacitors. *J. Electrochem. Soc.* **150,** A292–A300 (2003).
52. Sitte, W. Electrochemical cell for composition dependent measurements of chemical diffusion coefficients and ionic conductivities on mixed conductors and application to silver telluride at 160 °C. *Solid State Ion.* **59,** 117–124 (1993).

ARTICLE

doi:10.1038/nature18959

Tempo and mode of genome evolution in a 50,000-generation experiment

Olivier Tenaillon[1]*, Jeffrey E. Barrick[2,3]*, Noah Ribeck[3,4], Daniel E. Deatherage[2], Jeffrey L. Blanchard[5], Aurko Dasgupta[2]†, Gabriel C. Wu[2], Sébastien Wielgoss[6,7], Stéphane Cruveiller[8], Claudine Médigue[8], Dominique Schneider[7,9] & Richard E. Lenski[3,4]*

Adaptation by natural selection depends on the rates, effects and interactions of many mutations, making it difficult to determine what proportion of mutations in an evolving lineage are beneficial. Here we analysed 264 complete genomes from 12 *Escherichia coli* populations to characterize their dynamics over 50,000 generations. The populations that retained the ancestral mutation rate support a model in which most fixed mutations are beneficial, the fraction of beneficial mutations declines as fitness rises, and neutral mutations accumulate at a constant rate. We also compared these populations to mutation-accumulation lines evolved under a bottlenecking regime that minimizes selection. Nonsynonymous mutations, intergenic mutations, insertions and deletions are overrepresented in the long-term populations, further supporting the inference that most mutations that reached high frequency were favoured by selection. These results illuminate the shifting balance of forces that govern genome evolution in populations adapting to a new environment.

Comparative genomic studies have identified the molecular basis of adaptations including lactase permanence in humans[1], domestication of plants[2] and animals[3], and pathogenicity in bacteria[4]. Nevertheless, it is difficult to determine more generally what fraction of new mutations in an evolving lineage are beneficial. Answering this question is important for modelling sequence changes used in phylogenetic methods[5] and would inform debate about adaptive and non-adaptive modes of genome evolution[6,7].

The combination of experimental evolution and genome sequencing provides a way forward that has been used with viruses, bacteria, yeast and flies[8–13]. In a study of bacteria, the diversity of mutations involved in adaptation to high-temperature stress was studied by sequencing >100 lineages after a 2,000-generation experiment[10]. In another study, sequencing a series of clones from one population over 40,000 generations showed the trajectory of genome evolution[9]. However, a short-term experiment reveals only the early steps of adaptation, and it is difficult to distinguish adaptive 'driver' and non-adaptive 'passenger' mutations when only one population is examined. Beneficial mutations can also be identified by lineage tracking[14] and genetic reconstruction[15] experiments, but these approaches become impractical after an initial selective sweep or when mutations become too numerous over time, respectively.

To overcome these limitations, we analysed complete genomes of 264 clones from 12 populations across 50,000 generations of the long-term evolution experiment (LTEE) with *E. coli*[16,17]. These populations have evolved in a defined medium with scarce resources since 1988. Mean fitness measured in competition with their ancestor increased by ~70% in that time[17]. The LTEE is a model system for studying many fundamental evolutionary questions[9,15–23].

Genome-wide mutations and hypermutability

We sequenced the genomes of two clones from each population after 500, 1,000, 1,500, 2,000, 5,000, 10,000, 15,000, 20,000, 30,000, 40,000 and 50,000 generations using the Illumina platform (Supplementary Data 1). We called mutations, including structural variants, using the *breseq* pipeline[24,25]. In total, we found 14,572 point mutations; 500 insertions of insertion sequence (IS) elements; 726 deletions and 1,132 insertions each ≤ 50 base pairs (bp) (small indels); and 267 deletions and 45 duplications each >50 bp (large indels). After 50,000 generations, average genome length declined by 63 kb (~1.4%) relative to the ancestor (Extended Data Fig. 1). Mutations were not distributed uniformly across the populations. Instead, six populations (Ara−1, Ara−2, Ara−3, Ara−4, Ara+3 and Ara+6) had 96.5% of the point mutations, having evolved hypermutable phenotypes caused by mutations that affect DNA repair or removal of oxidized nucleotides[18,20]. Figure 1a shows the trajectories for the total mutations in all 12 populations; Fig. 1b is rescaled for better resolution of those that did not become point-mutation mutators. Hypermutability tended to decline over time as the load of deleterious mutations favoured antimutator alleles[20]. All four populations that were hypermutable at 10,000 generations accumulated synonymous substitutions (a proxy for the underlying point-mutation rate) between generations 40,000 and 50,000 at much lower rates than from 10,000 to 20,000 generations (Extended Data Fig. 2).

Increased numbers of IS elements can also cause hypermutability[26], with higher rates not only of transpositions but also deletions and duplications through homologous recombination. In population Ara+1, 31.8% of all mutations up to 50,000 generations were IS*150* insertions, compared with 12.3% for the other populations that never evolved elevated point-mutation rates. This mode of hypermutability arose early in Ara+1; IS*150* insertions are overrepresented in each Ara+1 clone from 5,000 generations onwards when compared individually to all other non-mutator clones from the same generation (Fisher's exact test with Bonferroni correction, $P < 0.05$). Two clones from other populations were also IS*150* hypermutators by this test: 38.7% of the mutations in

[1]IAME, UMR 1137, INSERM, Université Paris Diderot, Sorbonne Paris Cité, F-75018 Paris, France. [2]Department of Molecular Biosciences, Institute for Cellular and Molecular Biology, Center for Systems and Synthetic Biology, Center for Computational Biology and Bioinformatics, The University of Texas at Austin, Austin, Texas 78712, USA. [3]BEACON Center for the Study of Evolution in Action, Michigan State University, East Lansing, Michigan 48824, USA. [4]Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, Michigan 48824, USA. [5]Department of Biology, University of Massachusetts, Amherst, Massachusetts 01003, USA. [6]Institute of Integrative Biology, ETH Zürich, Universitätstrasse 16, Zürich 8092, Switzerland. [7]Université Grenoble Alpes, Laboratoire Technologies de l'Ingénierie Médicale et de la Complexité — Informatique, Mathématiques et Applications (TIMC-IMAG), F-38000 Grenoble, France. [8]UMR 8030, CNRS, Université Évry-Val-d'Essonne, CEA, Institut de Génomique, Laboratoire d'Analyses Bioinformatiques pour la Génomique et le Métabolisme, F-91000 Évry, France. [9]Centre National de la Recherche Scientifique, TIMC-IMAG, F-38000 Grenoble, France. †Present address: Department of Internal Medicine, Washington University School of Medicine, St Louis, Missouri 63110, USA.
*These authors contributed equally to this work.

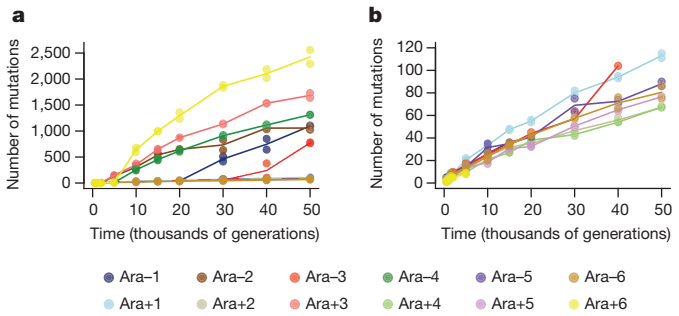


Figure 1 | Total number of mutations over time in the 12 LTEE populations. a, Total mutations in each population. **b**, Total mutations rescaled to reveal the trajectories for the six populations that did not become hypermutable for point mutations, and for the other six before they evolved hypermutability. Each symbol shows a sequenced genome; some points are hidden behind others. Each line passes through the average of the genomes from the same population and generation.

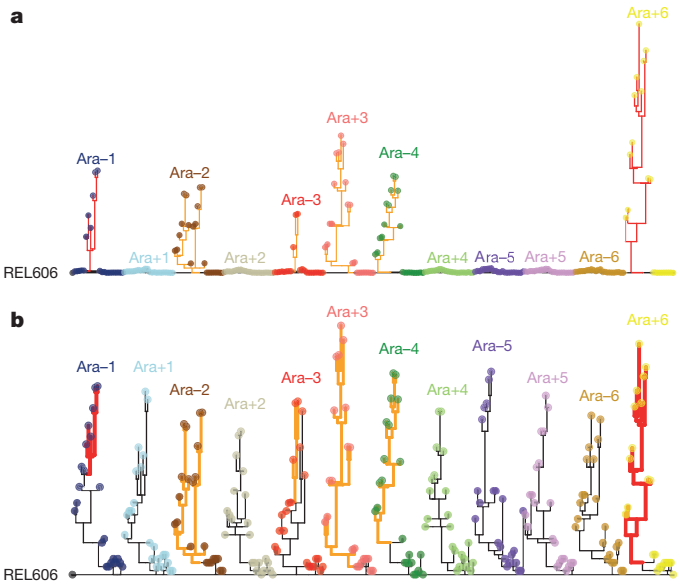


Figure 2 | Phylogenetic trees for LTEE populations. a, Phylogenies for 22 genomes from each population, based on point mutations. **b**, The same trees, except branches are rescaled as follows: branches for lineages with mismatch-repair defects are orange and shortened by a factor of 25; branches for *mutT* mutators are red and shortened by a factor of 50. Strain REL606 (on the left) is the ancestor. No early mutations are shared between any populations, confirming their independent evolution. Most populations have multiple basal lineages that reflect early diversification and extinction; some have deeply divergent lineages with sustained persistence, most notably Ara−2.

a 30,000-generation clone from Ara−5 and 31.7% of the mutations in a 40,000-generation clone from Ara−3 were IS*150* insertions. The aberrant Ara−5 clone shares only one mutation with other sequenced Ara−5 clones, indicating early divergence; it does not share point mutations with any other population, excluding cross-contamination. The emergence of these various mutator types shows that evolution can alter the production of genetic diversity[20,27], which in turn changes the tempo and mode of genome evolution.

Population phylogenies

Figure 2a shows phylogenetic trees constructed using point mutations for each population; Fig. 2b shows the trees with branches rescaled after mutators evolved. Some populations—including Ara−2, which became hypermutable early, and Ara−6, which never did—harbour lineages that coexisted for tens of thousands of generations. Some others—including Ara−4, which became hypermutable, and Ara+2, which did not—are more linear in structure, without deep branches among the sequenced clones. Deep branches were probably supported by the diversity-promoting effects of negative-frequency-dependent interactions, as shown in the Ara−2 population[22,23]. Sequencing whole-population samples would provide more detailed information on within-population diversity[11,12].

Dynamics of genome evolution

The accumulation of point mutations increased greatly in hypermutable populations[9,19,20], potentially overwhelming the genomic signature of adaptation. Although mutator lineages may experience higher rates of fitness improvement[17,27], the effect is usually small owing to clonal interference between competing beneficial mutations[28,29] and the increased load of deleterious mutations[20,30]. Therefore, beneficial mutations become harder to detect in a sea of unselected mutations in mutator lineages. To understand better the dynamic coupling between adaptation and genome evolution, we first analysed the populations that retained the ancestral mutation rate up to 50,000 generations and the others before they became point-mutation or IS*150* mutators.

It was previously found[17] that the mean-fitness trajectory of the LTEE is well described by a power-law relation, in which log fitness increases linearly with log time. Moreover, the power law accurately predicts fitness to 50,000 generations using data from only the first 5,000 generations. It was shown that a population-dynamical model that incorporates two phenomena known to be important in the LTEE—clonal interference[29,31] and diminishing-returns epistasis[15,29]—generates a power-law relation. This model in turn predicts that the number of beneficial mutations should increase with the square root of time[17]. However, not all mutations that accumulate are beneficial; neutral and nearly neutral mutations can spread by recurring mutation, random drift, and hitchhiking[32–34]. Selective sweeps will purge some neutral mutations but cause others to increase; overall, the expected number of neutral mutations should increase linearly with time[35].

To test these predictions, we fit three models to the trajectory for the total number of mutations in the non-mutator and premutator lineages:

$$m = at$$

$$m = b\sqrt{t}$$

$$m = at + b\sqrt{t}$$

where m is the number of mutations, t is time (generations), and a and b govern the genome-wide rates of accumulation of neutral and beneficial mutations, respectively (Fig. 3). (Extended Data Fig. 3 shows the models fit to each population separately.) Using the Akaike information criterion (AIC), the two-parameter model fits the data much better than those with only the linear ($\Delta\mathrm{AIC} = -77.7$) or square-root ($\Delta\mathrm{AIC} = -99.7$) terms. Because the one-parameter models are nested within the two-parameter model, we can also assess the significance of adding the second parameter; P values are 7.5×10^{-5} and 5.2×10^{-7} relative to the linear and square-root models, respectively. The trajectory for genome evolution thus shows signatures of both adaptive and non-adaptive changes. However, the model that predicts the square-root trajectory of beneficial substitutions makes various assumptions (for example, about the form of epistasis), and both the predicted and observed trajectories have statistical uncertainties. (Extended Data Fig. 4 shows the uncertainty in estimating a and b from the observed trajectory.) Therefore, we examined additional evidence to shed light on the proportion and identity of beneficial mutations.

Evidence for beneficial mutations

We sought to understand what proportion of the genomic changes in the non-mutator populations was adaptive, and how that proportion changed over time. One line of evidence derives from the expectation that synonymous substitutions—point mutations in protein-coding genes that do not affect the amino-acid sequence—are neutral and should therefore accumulate at a rate equal to the underlying

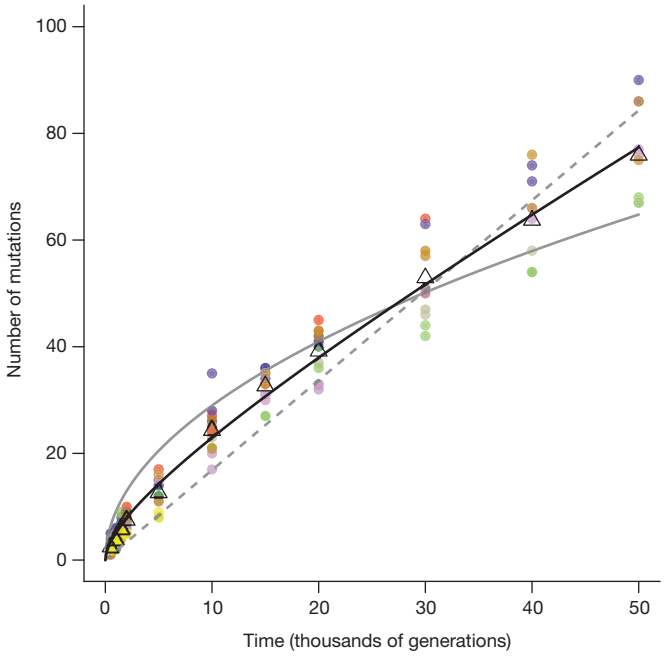


Figure 3 | Alternative models fit to the trajectory of genome evolution. Each symbol shows total mutations in a clone from five populations that never became mutators and seven before point mutation or IS*150* hypermutability evolved. Colours are the same as in Fig. 1; open triangles indicate grand means. Dashed grey line shows the best fit to the linear model, $m = at$. Solid grey curve shows the fit to the square-root model, $m = b\sqrt{t}$. Black curve is fit to the composite model, $m = at + b\sqrt{t}$, where $a = 0.000944$ and $b = 0.134856$. See text for statistical analysis.

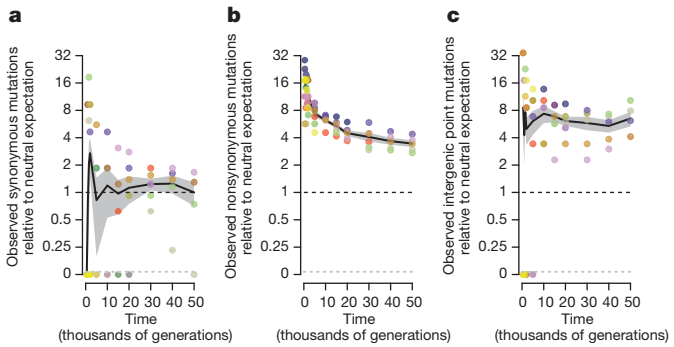


Figure 4 | Trajectories for synonymous, nonsynonymous and intergenic point mutations. a, Synonymous mutations, scaled so that the mean of five non-mutator populations (excluding point mutation and IS*150* hypermutators) is unity at 50,000 generations. **b**, Nonsynonymous mutations, scaled using the same rate as synonymous mutations after adjusting for sites at risk for both classes. **c**, Intergenic point mutations, scaled using the same rate as synonymous mutations after adjusting for sites at risk. Each symbol shows the mean for sequenced genomes from a non-mutator or premutator lineage. Colours are as in Fig. 1. Note the discontinuous scale; populations with zero mutations are plotted below. Black lines connect grand means; shading shows standard errors calculated from replicate populations.

mutation rate[20,35]. This expectation is not strictly true owing to selection on codon usage, RNA folding, and other effects, but it is generally thought that such selection is extremely weak, affects only a small fraction of sites at risk for synonymous mutations, or both[36,37]. We calculate whether nonsynonymous and intergenic point mutations are found in excess relative to synonymous mutations, given the number of sites at risk for each class. Figure 4a shows the number of synonymous mutations in non-mutator and premutator populations, scaled so the mean at 50,000 generations is unity. As expected, synonymous mutations accumulated at an approximately constant rate (Extended Data Fig. 5). Figure 4b shows the number of nonsynonymous mutations relative to the neutral expectation based on synonymous mutations. Nonsynonymous mutations accumulated ~17.1 times faster than synonymous ones during the first 500 generations and ~3.4 times faster over 50,000 generations. Nonsynonymous mutations continued to accumulate at over twice the rate of synonymous mutations in the later generations (Extended Data Fig. 6), implying that most nonsynonymous mutations that reached high frequency were beneficial even after so long in a constant environment. The same approach applied to intergenic point mutations (Fig. 4c) also reveals a large excess relative to synonymous mutations, although the number of events is smaller and the uncertainty greater. This result implicates adaptive changes in noncoding regions that presumably affect the binding sites for regulatory proteins[38–40].

Synonymous mutations provide an internal benchmark for nonsynonymous and intergenic point mutations. However, synonymous mutations are not directly informative for understanding how selection affects the accumulation of indels that comprise almost half the mutations in non-mutator clones at 50,000 generations (Extended Data Fig. 7). To estimate the proportion of beneficial changes for other types of mutation, we compare the LTEE and a mutation accumulation experiment (MAE) in which 15 lines were propagated via repeated single-cell bottlenecks[41]. Such bottlenecks eliminate the variation needed for natural selection, so that all types of mutations accumulate at the rates at which they happen, regardless of fitness effects, except for lethal or highly deleterious mutations that preclude cells from making colonies used to propagate lines[29]. MAE lines thus provide an external baseline for distinguishing beneficial and non-beneficial mutations. In fact, because more unselected mutations are deleterious than beneficial, MAE lines are expected to lose fitness over time, which they did (Extended Data Fig. 8).

To quantify the relative rates for all types of mutations in the absence of selection, we sequenced clones from the MAE lines after 550 daily bottlenecks (Supplementary Data 1). Consistent with the random accumulation of mutations, the number of nonsynonymous (including nonsense) mutations was similar to the expectation based on synonymous mutations (117 observed, 105.02 expected); the resulting ratio of 1.11 is well within the 95% confidence interval (0.70–1.50) obtained by a randomization test. Also, there was no among-line variation in total mutations ($\chi^2 = 5.46$, degrees of freedom (df) = 14, $P = 0.978$). We can therefore reasonably use the MAE lines to estimate relative rates of different types of mutations, with synonymous ones providing a benchmark largely free of selection in both experiments. For example, LTEE population Ara−1 had 21 nonsynonymous mutations at 20,000 generations and the expected number of synonymous mutations based on the average non-mutator population was 1.08 (Extended Data Fig. 5); the 15 MAE lines in total had 117 nonsynonymous and 39 synonymous mutations; thus, the ratio of observed mutations to the neutral expectation is (21/1.08)/(117/39) = 6.5. These ratios show that all major classes of mutations—including various indels—are substantially overrepresented in the LTEE relative to the MAE (Extended Data Fig. 9), implying that many mutations in each class were adaptive during the LTEE.

Parallel evolution at many gene loci

Parallel evolution occurs when similar changes arise independently in multiple lineages, and it is often used to discover putative targets of selection[4,8,10–13,21]. Genetic parallelism can be studied at the level of DNA sequence, affected genes, or integrated functions. Parallelism at the nucleotide level tends to be rare because different mutations in a gene often produce similar benefits[4,10–12,21], although there are exceptions[8]. Parallelism at a functional level requires detailed understanding that may be unavailable, and it is difficult to interpret when there are many mutations. We therefore examined parallelism at the gene level.

Table 1 | Protein-coding genes with the highest *G* scores

Gene	Length	Observed	Expected	*G*	Annotation
pykF	1,413	19	0.16	181	Pyruvate kinase
icIR	825	13	0.10	128	Transcriptional repressor, glyoxylate bypass
spoT	2,109	14	0.25	113	Stringent response
nadR	1,233	12	0.14	106	Bifunctional transcriptional repressor and NMN adenylyltransferase
hslU	1,332	11	0.15	94	Molecular chaperone and ATPase component of protease
yijC (also known as *fabR*)	705	7	0.08	62	Transcriptional repressor, fatty acid and phosphatidic acid pathway
topA	2,598	8	0.30	52	DNA topoisomerase I subunit
malT	2,706	8	0.31	52	Transcriptional activator, maltotriose-ATP-binding
mrdA	1,902	7	0.22	48	Transpeptidase in peptidoglycan synthesis
mreB	1,044	6	0.12	47	Longitudinal peptidoglycan synthesis
infB	2,673	7	0.31	44	Translation initiation factor IF-2
arcA	717	5	0.08	41	Response regulator in two-component system, anoxic redox control
argR	471	4	0.05	34	Repressor of arginine regulon
rplF	534	4	0.06	33	50S ribosomal subunit protein
mreC	1,104	4	0.13	28	Longitudinal peptidoglycan synthesis

Genes are ranked by *G* scores computed using observed independent nonsynonymous mutations relative to expected number given gene length (bp). Data are from populations with the ancestral point-mutation rate throughout and other populations before they evolved hypermutability.

We focused on lineages that retained the ancestral point-mutation rate (including clones from populations that later became hypermutable) because, as shown earlier, most mutations are drivers in those cases; we expect hypermutability to make the analysis less informative because many more mutations are passengers. We first calculated the expected number of nonsynonymous mutations for each single-copy protein-coding gene based on its length as a fraction of all such genes and the total number of nonsynonymous mutations in the relevant lineages (Supplementary Data 2). We computed *G* scores for goodness of fit between observed and expected values; the total score is 2,593.7. We compared that total with simulated data sets in which positions of mutations in the coding genome were randomized, and the observed total significantly exceeded the simulations (mean simulated $G = 1{,}933.7$, $Z = 25.5$, $P < 10^{-143}$). Fifty-seven genes had two or more mutations; these genes had 50.1% of the nonsynonymous mutations but constituted only 2.1% of the coding genome. (Only one gene had multiple synonymous changes.) Table 1 shows the 15 genes that contribute the most to the total *G* score. Several encode proteins with core metabolic or regulatory functions, including three involved in peptidoglycan synthesis.

We ran the same analysis for lineages that evolved hypermutability (Supplementary Data 3), and the randomization test indicates significant parallelism (G statistic $= 5{,}098.4$, mean simulated $G = 4{,}581.1$, $Z = 5.745$, $P < 10^{-8}$). As expected, however, the signal-to-noise ratio reflected in the significance level is much weaker than for the non-mutator lineages. Most genes with the highest scores in mutator lineages differ from those in non-mutators, in part because those genes often had beneficial mutations before hypermutability evolved.

Table 2 lists the 16 genes with the most deletions, duplications, insertions and intergenic point mutations in non-mutator lineages

Table 2 | Genes with the most mutations of other types

Genes	Mutations	Number	IS	MAE	Annotation
rbsD	Mostly large deletions	41	Yes	No	D-Ribose utilization; most deletions affect entire *rbs* operon
nupC	Various intergenic	19	Yes	Yes	Nucleoside transporter
iap	Mostly large indels	19	Yes	No	Alkaline-phosphatase isozyme conversion; most indels affect tens of adjacent genes including *rpoS*, which encodes stationary-phase σ factor
mokB	Various indels	17	Yes	Yes	Enables *hokB* toxin expression
yhgI/gntT	Intergenic point mutations	16	No	No	Gluconate transport
mokC	Various indels	15	Yes	Yes	Enables *hokC* toxin expression
ybcU (also known as *borD*)	Large indels	14	Yes	No	Indels affect this and adjacent remnants of DLP12 prophage
ECB_02013	Various indels	14	No	Yes	Indels affect this and adjacent remnants of P2-like prophage
ECB_02816 (also known as *kpsD*)	Various indels	14	Yes	No	Polysialic-acid transport protein precursor
acs/nrfA	Various intergenic	14	No	No	Acetyl-CoA synthase; nitrite reductase
hokE	Large indels	12	Yes	No	Toxin in plasmid-derived toxin–antitoxin system; most indels affect several adjacent genes involved in iron acquisition
ybeB/phpB	Various intergenic	11	Yes	No	Unknown functions, but adjacent to genes involved in cell-wall synthesis
ydiJ/ydiK	Various intergenic	11	No	No	Predicted FAD-linked oxidoreductase; putative inner membrane protein
ldrC	Various indels	10	Yes	Yes	Small toxic polypeptide
menC	IS insertions	10	Yes	Yes	Menaquinone biosynthesis
fimA	Mostly IS insertions	10	Yes	No	Component of fimbrial complex

Genes are ranked by total mutations excluding nonsynonymous and synonymous point mutations. When two genes are separated by a solidus, the affected sequence includes the intergenic region between them. IS column indicates whether the majority of mutations involve IS elements. MAE column indicates whether the same or nearly identical mutations occurred in one or more MAE lines. Data are from populations with the ancestral point-mutation rate throughout and others before they evolved hypermutability.

(Supplementary Data 2). For mutations that impact multiple genes, we show the most frequently affected gene (or adjacent pair when most events are intergenic). In 12 cases, the majority of the mutations were mediated by IS elements; these include insertions as well as deletions and duplications that appear to involve homologous recombination. In six cases (five with IS insertions), the same or nearly identical mutations occurred in one or more MAE lines, suggesting mutational hotspots. These changes may indicate high-frequency events, but recall that IS insertions and large indels are enriched in the LTEE relative to the MAE (Extended Data Fig. 9), implying that many are also beneficial. Indeed, the IS-mediated *rbsD* deletions occur at a high rate and are beneficial in the LTEE environment[42], and some IS-mediated mutations appear to be beneficial in other studies as well[43,44].

The parallelisms involving nonsynonymous substitutions and other mutations in the LTEE, coupled with their high rates of accumulation relative to the MAE, indicate that many observed mutations were drivers of adaptation. For indels, however, the specific target genes are difficult to identify owing to the multiplicity of genes affected and the potentially confounding effect of mutational hotspots.

Discussion

Adaptation by natural selection sits at the heart of phenotypic evolution. However, the random processes of spontaneous mutation and genetic drift often overwhelm and obscure genomic signatures of adaptation. We overcame this difficulty by analysing genomes from 12 bacterial populations that evolved for 50,000 generations under identical culture conditions. Even so, six populations evolved hypermutable phenotypes that increased point-mutation rates ~100-fold, and another evolved hypermutability caused by a transposable element. By focusing on populations that retained the ancestral mutation rate, we identified several key features of the tempo and mode of their genome evolution. First, a population-genetic model with two terms—one for beneficial drivers, the other for neutral hitchhikers—fits the dynamics much better than models without both terms. Second, the great majority of mutations observed during the early generations were beneficial drivers. Third, the proportion of observed mutations that were beneficial declined over time but remained substantial even after 50,000 generations. The second and third findings follow from the population-genetic model. Both are also strongly supported by the excess of nonsynonymous to synonymous substitutions in the LTEE and by the excess of several classes of mutations, including indels, in comparison to mutation-accumulation lines. Fourth, there was strong gene-level parallel evolution across the replicate LTEE populations.

Our analyses also show a contrast between the contributions of beneficial mutations to molecular evolution and to the fitness trajectory in a stable environment. In particular, beneficial mutations continued to constitute a large fraction of genetic changes throughout the 50,000 generations of the LTEE, whereas the resulting fitness gains were only a few per cent in the last 10,000 generations[17]. Beneficial mutations with very small selection coefficients are nonetheless visible to natural selection[17]. Hence, adaptation can remain a major driver of molecular evolution long after an environmental shift. Our experimental results thus support a selectionist view of molecular evolution, complementing indirect evidence based on comparative genomics in bacteria, *Drosophila* and humans[45–47]. Of course, the LTEE may differ from many natural populations in important respects including its low mutation rate, the absence of sex or horizontal gene transfer, and a stable environment. As we showed, high mutation rates tend to obscure the role of selection in molecular evolution. The effects of horizontal gene transfer[48] and variable environments[49,50] on the dynamic coupling of genomic and adaptive evolution should also be examined further. Long-term experiments with microorganisms provide opportunities for rigorous analyses of these issues.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 24 January; accepted 23 June 2016.
Published online 1 August 2016.

1. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74,** 1111–1120 (2004).
2. Hufford, M. B. *et al.* Comparative population genomics of maize domestication and improvement. *Nature Genet.* **44,** 808–811 (2012).
3. Vonholdt, B. M. *et al.* Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature* **464,** 898–902 (2010).
4. Lieberman, T. D. *et al.* Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nature Genet.* **43,** 1275–1280 (2011).
5. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24,** 1586–1591 (2007).
6. Lynch, M. & Conery, J. S. The origins of genome complexity. *Science* **302,** 1401–1404 (2003).
7. Whitney, K. D. & Garland, T. Jr Did genetic drift drive increases in genome complexity? *PLoS Genet.* **6,** e1001080 (2010).
8. Wichman, H. A., Badgett, M. R., Scott, L. A., Boulianne, C. M. & Bull, J. J. Different trajectories of parallel evolution during viral adaptation. *Science* **285,** 422–424 (1999).
9. Barrick, J. E. *et al.* Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* **461,** 1243–1247 (2009).
10. Tenaillon, O. *et al.* The molecular diversity of adaptive convergence. *Science* **335,** 457–461 (2012).
11. Lang, G. I. *et al.* Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature* **500,** 571–574 (2013).
12. Kvitek, D. J. & Sherlock, G. Whole genome, whole population sequencing reveals that loss of signaling networks is the major adaptive strategy in a constant environment. *PLoS Genet.* **9,** e1003972 (2013).
13. Burke, M. K. *et al.* Genome-wide analysis of a long-term evolution experiment with *Drosophila*. *Nature* **467,** 587–590 (2010).
14. Levy, S. F. *et al.* Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature* **519,** 181–186 (2015).
15. Khan, A. I., Dinh, D. M., Schneider, D., Lenski, R. E. & Cooper, T. F. Negative epistasis between beneficial mutations in an evolving bacterial population. *Science* **332,** 1193–1196 (2011).
16. Lenski, R. E., Rose, M. R., Simpson, S. C. & Tadler, S. C. Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2000 generations. *Am. Nat.* **138,** 1315–1341 (1991).
17. Wiser, M. J., Ribeck, N. & Lenski, R. E. Long-term dynamics of adaptation in asexual populations. *Science* **342,** 1364–1367 (2013).
18. Sniegowski, P. D., Gerrish, P. J. & Lenski, R. E. Evolution of high mutation rates in experimental populations of *E. coli*. *Nature* **387,** 703–705 (1997).
19. Blount, Z. D., Barrick, J. E., Davidson, C. J. & Lenski, R. E. Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature* **489,** 513–518 (2012).
20. Wielgoss, S. *et al.* Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proc. Natl Acad. Sci. USA* **110,** 222–227 (2013).
21. Woods, R., Schneider, D., Winkworth, C. L., Riley, M. A. & Lenski, R. E. Tests of parallel molecular evolution in a long-term experiment with *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **103,** 9107–9112 (2006).
22. Rozen, D. E. & Lenski, R. E. Long-term experimental evolution in *Escherichia coli*. VIII. Dynamics of a balanced polymorphism. *Am. Nat.* **155,** 24–35 (2000).
23. Plucain, J. *et al.* Epistasis and allele specificity in the emergence of a stable polymorphism in *Escherichia coli*. *Science* **343,** 1366–1369 (2014).
24. Deatherage, D. E. & Barrick, J. E. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using *breseq*. *Methods Mol. Biol.* **1151,** 165–188 (2014).
25. Barrick, J. E. *et al.* Identifying structural variation in haploid microbial genomes from short-read resequencing data using *breseq*. *BMC Genomics* **15,** 1039 (2014).
26. Chao, L., Vargas, C., Spear, B. B. & Cox, E. C. Transposable elements as mutator genes in evolution. *Nature* **303,** 633–635 (1983).
27. Tenaillon, O., Taddei, F., Radmian, M. & Matic, I. Second-order selection in bacterial evolution: selection acting on mutation and recombination rates in the course of adaptation. *Res. Microbiol.* **152,** 11–16 (2001).
28. Gerrish, P. J. & Lenski, R. E. The fate of competing beneficial mutations in an asexual population. *Genetica* **102–103,** 127–144 (1998).
29. Barrick, J. E. & Lenski, R. E. Genome dynamics during experimental evolution. *Nature Rev. Genet.* **14,** 827–839 (2013).
30. Good, B. H. & Desai, M. M. Deleterious passengers in adapting populations. *Genetics* **198,** 1183–1208 (2014).
31. Maddamsetti, R., Lenski, R. E. & Barrick, J. E. Adaptation, clonal interference, and frequency-dependent interactions in a long-term evolution experiment with *Escherichia coli*. *Genetics* **200,** 619–631 (2015).
32. Gillespie, J. H. Genetic drift in an infinite population. The pseudohitchhiking model. *Genetics* **155,** 909–919 (2000).

33. Neher, R. A. & Shraiman, B. I. Genetic draft and quasi-neutrality in large facultatively sexual populations. *Genetics* **188,** 975–996 (2011).
34. Kosheleva, K. & Desai, M. M. The dynamics of genetic draft in rapidly adapting populations. *Genetics* **195,** 1007–1025 (2013).
35. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, 1983).
36. Sharp, P. M., Emery, L. R. & Zeng, K. Forces that influence the evolution of codon bias. *Phil. Trans. R. Soc. Lond. B* **365,** 1203–1212 (2010).
37. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nature Rev. Genet.* **12,** 32–42 (2011).
38. Stern, D. L. Evolutionary developmental biology and the problem of variation. *Evolution* **54,** 1079–1091 (2000).
39. Carroll, S. B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134,** 25–36 (2008).
40. Oren, Y. *et al.* Transfer of noncoding DNA drives regulatory rewiring in bacteria. *Proc. Natl Acad. Sci. USA* **111,** 16112–16117 (2014).
41. Kibota, T. T. & Lynch, M. Estimate of the genomic mutation rate deleterious to overall fitness in *E. coli*. *Nature* **381,** 694–696 (1996).
42. Cooper, V. S., Schneider, D., Blot, M. & Lenski, R. E. Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli* B. *J. Bacteriol*. **183,** 2834–2841 (2001).
43. Miskinyte, M. *et al.* The genetic basis of *Escherichia coli* pathoadaptation to macrophages. *PLOS Pathog.* **9,** e1003802 (2013).
44. Wielgoss, S., Bergmiller, T., Bischofberger, A. M. & Hall, A. R. Adaptation to parasites and costs of parasite resistance in mutator and nonmutator bacteria. *Mol. Biol. Evol.* **33,** 770–782 (2016).
45. Charlesworth, J. & Eyre-Walker, A. The rate of adaptive evolution in enteric bacteria. *Mol. Biol. Evol.* **23,** 1348–1356 (2006).
46. Sawyer, S. A., Kulathinal, R. J., Bustamante, C. D. & Hartl, D. L. Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. *J. Mol. Evol.* **57** (suppl. 1), S154–S164 (2003).
47. Bustamante, C. D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* **437,** 1153–1157 (2005).
48. Cooper, T. F. Recombination speeds adaptation by reducing competition between beneficial mutations in populations of *Escherichia coli*. *PLoS Biol.* **5,** e225 (2007).
49. Satterwhite, R. S. & Cooper, T. F. Constraints on adaptation of *Escherichia coli* to mixed-resource environments increase over time. *Evolution* **69,** 2067–2078 (2015).
50. Paterson, S. *et al.* Antagonistic coevolution accelerates molecular evolution. *Nature* **464,** 275–278 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank N. Hajela for assistance, R. Maddamsetti and Z. Blount for discussions, and M. Lynch for starting the MAE lines. This research was supported by the US National Science Foundation (DEB-1451740 to R.E.L.), BEACON Center for the Study of Evolution in Action (DBI-0939454), European Research Council (FP7 grant 310944 to O.T.), European Union (FP7 grant 610427 to D.S.), French National Funding Agency (ANR-08-GENM-023-001 to D.S., O.T. and C.M.), French CNRS International Associated Laboratory (to D.S. and R.E.L.), and US National Institutes of Health (R00-GM087550 to J.E.B.). D.E.D. was supported by a traineeship from the Cancer Prevention and Research Institute of Texas. We acknowledge the use of high-performance computing resources at the Texas Advanced Computing Center.

Author Contributions O.T., J.E.B., D.S. and R.E.L. conceived the project; R.E.L. and J.L.B. provided strains; O.T., J.E.B., D.E.D., A.D., G.C.W., S.W., S.C. and C.M. analysed genomes and generated other data; N.R. developed theory; R.E.L., O.T. and J.E.B. wrote the paper. All authors approved the submitted version.

Author Information All sequencing data sets are available in the NCBI BioProject database under accession number PRJNA294072. The *breseq* analysis pipeline is available at GitHub (http://github.com/barricklab/breseq). Other analysis scripts are available at the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.6226d). R.E.L. will make strains available to qualified recipients, subject to a material transfer agreement. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.E.L. (lenski@msu.edu).

Reviewer Information *Nature* thanks M. Desai, G. Sherlock and C. Zeyl for their contribution to the peer review of this work.

METHODS

Long-term evolution experiment. The LTEE has 12 populations founded from two almost identical strains of *Escherichia coli*. Six populations, designated Ara−1 to Ara−6, started from REL606, a descendant of the B strain of Luria and Delbrück[51–53]. The other six, Ara+1 to Ara+6, derive from REL607, which differs from REL606 by point mutations in *araA* and *recD*. The mutation in *araA* was selected before starting the LTEE; it confers the ability to grow on L-arabinose, which provides a marker in competition assays used to measure fitness[16,17]. The *recD* mutation arose inadvertently before starting the LTEE. The LTEE began in 1988, and the populations have been propagated (with occasional interruptions) at 37 °C by daily 100-fold dilutions in 10 ml Davis minimal medium with 25 μg/ml glucose (http://lenski.mmg.msu.edu/ecoli/dm25liquid.html). The regrowth allows ~6.67 generations per day; the population size fluctuates between $\sim 3 \times 10^6$ and $\sim 3 \times 10^8$ cells except in population Ara−3, which has had a population size several times larger since ~33,000 generations, when cells gained the ability to consume the citrate that is also present in the medium[19,54]. Whole-population samples are taken every 75th transfer (500 generations) and stored with glycerol as a cryoprotectant at −80 °C, where they are available for later analysis. Here we analysed the genomes of two clones sampled from each population at 500, 1,000, 1,500, 2,000, 5,000, 10,000, 15,000, 20,000, 30,000, 40,000 and 50,000 generations (Supplementary Data 1). We deliberately included clones from the deeply diverged lineages in population Ara−2 from 20,000 generations onwards and both the majority Cit^+ lineage and the minority Cit^- lineage in population Ara−3 at generation 40,000. This sampling scheme does not affect inferences about the rates and patterns of genome evolution because both populations were hypermutable at these time points and thus excluded from the main analyses. These clones were included to illustrate diversity within populations, although we also found previously unknown cases of divergent lineages. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded during experiments and outcome assessment.

Mutation-accumulation experiment. The 15 MAE lines analysed here started from strain REL1207, which is an Ara^+ mutant of a clone sampled from LTEE population Ara−1 at 2,000 generations. REL1207 differs from REL606 by a total of eight mutations, including one in *araA* that confers the Ara^+ marker phenotype. Each line was propagated through 550 single-cell bottlenecks by picking a colony at random from a Davis minimal agar plate with glucose at 200 μg/ml and streaking the cells onto a fresh plate. Given ~25 cell doublings to produce a typical colony[41], the 550 cycles represent ~13,750 generations. The bottlenecks imposed by this procedure eliminate the genetic variation that fuels adaptation by natural selection; as a consequence, mutations accumulate at rates that depend on their underlying mutation rate but not their fitness effects, except for highly deleterious mutations that preclude sufficient growth to form a colony[29]. Because more mutations are deleterious than are beneficial, fitness declined under this regime (Extended Data Fig. 8). The 15 sequenced clonal isolates, each from a different MAE line, are JEB807–JEB821 (Supplementary Data 1). None of the lineages became hypermutable based on their mutational signatures and the absence of significant heterogeneity in the total mutations accumulated (see main text). However, the mean per-generation rate at which synonymous mutations arose was ~3.5-fold higher in the MAE lines than in the five LTEE populations that remained non-mutators for all 50,000 generations (Supplementary Data 4; $t_s = 3.0755$, $P = 0.0065$). This difference may reflect the different conditions in liquid and agar media, including the glucose concentration and local cell density, which might affect the reactive oxygen species that cells experience. The comparisons between the LTEE and MAE (Extended Data Fig. 9) would change if the underlying rates of the various types of mutation responded disproportionately to the different conditions in the MAE. That possibility seems implausible for the different classes of point mutation (Extended Data Fig. 9a, b), and the differences would have to be substantially larger than the different rates of synonymous mutations to produce the excess IS*150* insertions (Extended Data Fig. 9c) and large indels (Extended Data Fig. 9f) observed in the LTEE relative to the MAE.

Genome sequencing. Frozen samples from the LTEE and MAE were revived via overnight growth at 37 °C in either LB or Davis minimal medium supplemented with 1,000 μg/ml glucose. Genomic DNA was isolated from each culture using the Qiagen Genomic-tip 100/G kit or equivalent. The DNA samples were sequenced at Genoscope or Integragen SA (Évry, France), the Michigan State University Research Technology Support Facility (East Lansing, USA), or the University of Texas at Austin Genome Sequencing and Analysis Facility (Austin, USA). Illumina Genome Analyzer and HiSeq instruments were used to generate single-end or paired-end reads ranging in length from 35 to 150 bases according to standard procedures, with median coverage of 80-fold and 95-fold for the 264 LTEE and 15 MAE clones, respectively (Supplementary Data 1). Of the 264 LTEE genomes in this study, 40 were previously analysed in other studies[9,19,20,55–57]. Supplementary Data 4 shows the number of every type of mutation inferred after performing the analyses described below on each of the LTEE and MAE genomes used in this study.

Mutation calling. We used *breseq* (versions 0.26.0 to 0.27.0) to predict both single-nucleotide and structural differences[24,25] based on how the Illumina reads for each sample mapped to the genome sequence of *E. coli* B REL606 (GenBank accession NC_012967.1)[52]. We counted and classified mutations using an updated version of the REL606 reference genome with improved feature annotations. The updated genome file (in both GenBank and GFF3 formats) and lists of predicted mutations in each evolved genome (in the Genome Diff format described in an appendix to the *breseq* manual) are freely available online (http://github.com/barricklab/LTEE-Ecoli).

Most types of single-step mutations, including large deletions and transposition events leading to copies of IS elements at new positions in the genome, are directly predicted by *breseq* when they occur in non-repetitive genomic regions. The initial lists of predicted mutations were curated and refined as previously described[24]. Briefly, complex mutations involving multiple steps (such as a new IS insertion followed by a flanking deletion) and structural mutations that overlap repetitive regions of the genome were manually resolved from unassigned new junction and missing coverage evidence in the *breseq* output. Large duplications and amplifications were detected by examining the coverage depth of mapped reads across the reference genome and comparing this information with the positions of repeat sequences and unassigned junctions. Owing to limitations of short-read DNA sequencing data, we could not fully predict point mutations and indels of one to a few base pairs within repeat regions (for example, IS elements) or gene conversions, in which intragenomic recombination between nearly identical copies of a large repeat region (for example, the seven copies of the rRNA operon) converts a minor variation in one copy to match exactly the sequence of another copy. Instead, all such genetic changes in repetitive regions of the genome were uniformly ignored in downstream analyses, as described later.

To validate the final lists of mutations predicted in each clone, we applied these changes to the ancestral REL606 sequence and used *breseq* to compare the Illumina reads against this simulated evolved genome to verify there were no further, unexplained discrepancies. This step of applying mutations to the reference genome was also used to estimate the final genome size of each evolved clone, with the assumption that new IS insertions were of the most common size for that IS element in the reference genome.

For 6 of the 264 LTEE samples, there was evidence of non-clonality in the sequence data. Some samples appeared to be mixtures of two very closely related clones that shared nearly all mutations but had one to several mutations specific to each type, together adding to a frequency of 100% (for example, sets of mutations at frequencies of 35% and 65%). This situation might result from inadvertently sampling two adjacent colonies on an agar plate when picking clones from an LTEE population. In other cases, only one or two mutations were found at an intermediate frequency. This type of heterogeneity might arise from strong selection favouring new mutations during colony outgrowth, subculturing and revival of samples before DNA extraction, as these conditions differ from the LTEE. In each case, we reconstructed the major genotype in the sample, as noted in Supplementary Data 1.

We also ignored putative genome variation associated with a cryptic 186-like prophage element (REL606 genome coordinates 880528–904682). In ten of the LTEE populations, we observed clones with increased read-coverage depth of this region and reads spanning a new sequence junction consistent with either tandem head-to-tail amplifications of this region or the production of circular DNA molecules joined at these exact nucleotides. The changes in the apparent copy number of this region often deviated from the integer values expected for a stable duplication or amplification. The prophage-related changes in coverage appeared most often in genomes isolated from 2,000 generations or earlier in the LTEE. There is no evidence of infective phage production in the LTEE, but it is possible that replication of DNA encoding a defective phage occurs stochastically at some low level in the ancestral strain REL606 or that production of this DNA is induced by stress when culturing samples for DNA isolation.

Phylogenetic consistency. Owing to the long duration of the LTEE and the evolution of mutators in several lineages, some mutations may be hidden or initially grouped with other mutations into a single change when comparing a late-generation evolved genome with the ancestral genome. For example, a point mutation might occur early in the experiment and then the region containing that mutation is later deleted. Similarly, the deletion of one base early and the subsequent deletion of an adjacent base would be called as a single two-base deletion in later samples. To obtain more accurate counts in light of these issues, we used each population's inferred phylogeny to split or add mutations, as appropriate, so that the mutation list for each clone reflects the most parsimonious set of mutational steps between that clone and its ancestor. Specifically, we chose histories with the fewest total mutations, the fewest mutations on early branches (in case of ties), and the fewest

total nucleotide changes summed over all mutations. Because this procedure is conservative in adding mutations to achieve phylogenetic consistency, it might underestimate the number of mutations on branches leading to an evolved genome when intermediate states are not resolved by the relationships of the sequenced clones.

Final mutation lists. We performed two final filtering steps to enable the sets of mutations to be uniformly compared across all genomes. In doing so, we classified as 'small mutations' all single-nucleotide substitutions, insertions and deletions of 20 or fewer bp, substitutions replacing 20 or fewer bp in the reference genome with 20 or fewer other bp, and all simple sequence repeat (SSR) mutations regardless of their size. SSR mutations add or remove one or more copies of a tandem-repeat unit consisting of one or a few bp. We defined SSR mutations as containing at least two copies of the repeat unit and having a total length of at least five bp when including all copies of the tandem repeat in the reference genome. For example, the genetic changes GGGGG→GGGG, TATATA→TATATATA and TACGTTACGT→TACGT would all be classified as SSR mutations, but GGGG→GGGGG, TATA→TATATA and TACGT→TACGTTACGT would not. All other genomic changes were considered 'large mutations' for purposes of filtering.

The ability to call small mutations located in repetitive regions of the genome is dependent on read length, so we removed all such mutations in regions where it would be a problem to uniformly detect them from the mutation lists before further analyses. To do this, we enumerated all regions of ≥ 20 bp that had an exact match elsewhere in the genome of the ancestral strain REL606 using MUMmer v.3.23 (ref. 58). We then merged regions from this list that were separated by five or fewer bp. All resulting regions that were now ≥ 35 bp were included in a list of masked genomic intervals. We also added to this list a hypervariable SSR consisting of seven copies of a tetranucleotide sequence that could not be reliably called in data sets with short reads (coordinates 2103889–2103919). Any small mutations contained in these masked regions were excluded from all downstream analyses.

Finally, we flagged all nucleotide substitutions or small indels occurring within 20 bp of the end of an IS element. The sequences directly adjacent to IS elements appear to experience an unusually high mutation rate, possibly due to frequent transposase cleavage and DNA repair. Mutations at these IS-adjacent sites probably have no effect on cellular phenotypes and fitness. We excluded them from the final lists of mutations used in all further analyses because they could bias the inferred mutational spectra and rates.

Phylogenetic analyses. To produce the phylogenetic trees shown in Fig. 2, we used the point mutations associated with each clone. A minimum-evolution tree was built using the Jukes–Cantor one-parameter model[59]. We used this model for two reasons. First, the mutator lineages had very different mutational spectra from the non-mutators[9,20,55,57]. Second, many mutations seen in non-mutator lineages were under positive selection, and so it is appropriate to give the mutations equal weight and not, for instance, reduce the importance of transitions relative to transversions. The trees were plotted with the R package APE[60]. The composite tree has the star-like structure expected for independent evolution of the populations. Therefore, trees were made separately for each population and then combined in Fig. 2, which allowed multiple basal branches to be placed with the appropriate populations.

Parallel evolution in non-mutator lineages. For genomes that did not come from point-mutation hypermutator lineages (Supplementary Data 1), we examined the extent of parallelism at the gene level in two ways. The first approach was based only on nonsynonymous mutations, because it is straightforward to quantify the overall extent of parallelism, determine the statistical significance of the parallelism, and rank genes based on their contributions to the significance. For each protein-coding gene i, we know its length, L_i, and the number of independent nonsynonymous mutations observed in that gene across all clones from non-mutator and premutator lineages, N_i. We summed the lengths and relevant mutations over all single-copy protein-coding genes in the ancestral genome to obtain L_{tot} (3,920,306) and N_{tot} (457, including two mutations that each affected overlapping reading frames), respectively. We computed the expected number of mutations in each gene, E_i, as follows:

$$E_i = N_{tot} (L_i/L_{tot})$$

We then computed a G_i score for each gene for which $N_i > 0$ as follows:

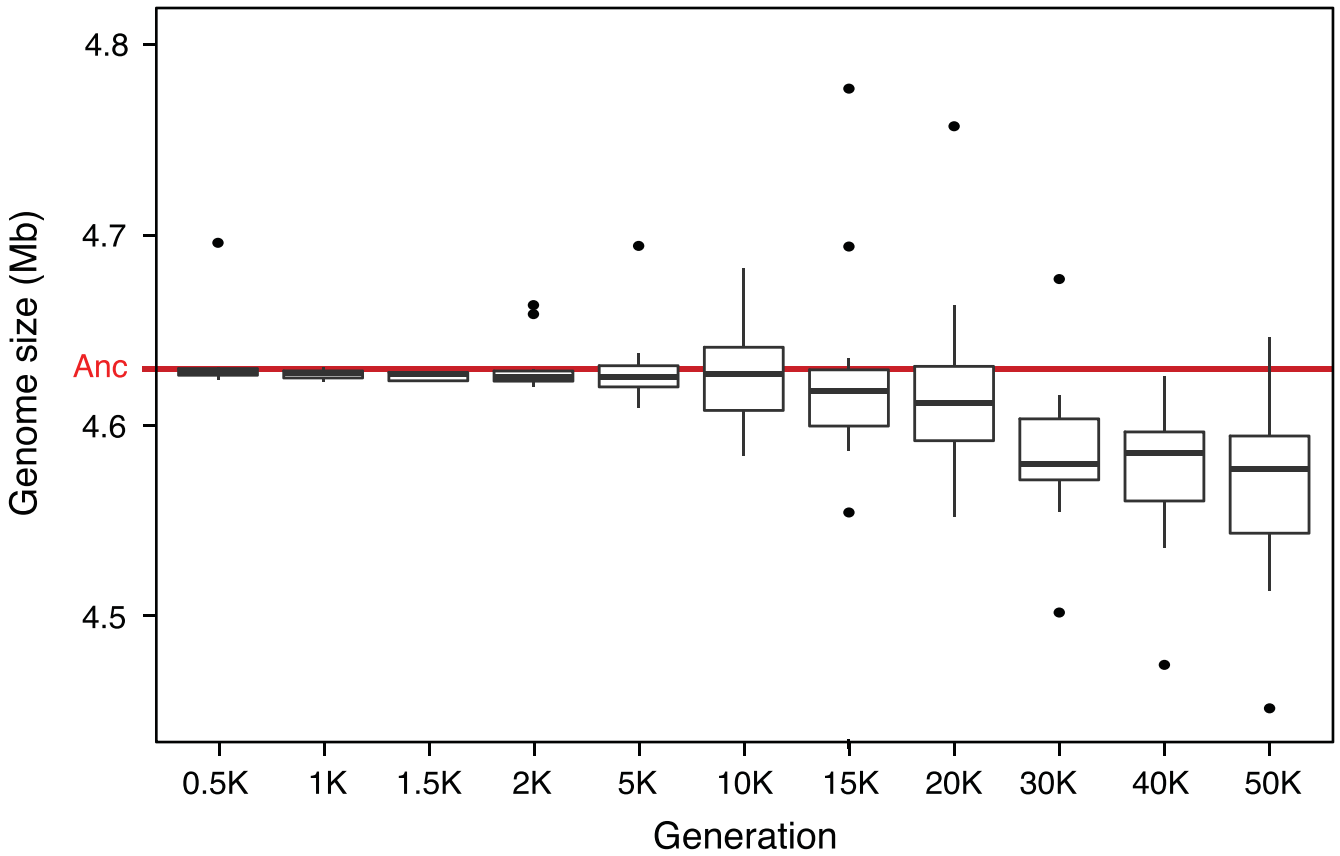
$$G_i = 2N_i \log_e(N_i/E_i)$$

We set $G_i = 0$ for those genes for which $N_i = 0$. This analysis ignores variability among genes in the proportion of sites at risk for nonsynonymous mutations. However, such differences are small and should hardly affect the analysis. The total G statistic equals the sum of the scores over all genes. To compute the expected G statistic under the null hypothesis of a random distribution of mutations, we generated 1,000 simulated data sets in which N_{tot} mutations were randomly placed throughout the coding genome. We computed the total G statistic for each simulated data set, and we calculated its mean and standard deviation across the 1,000 simulations. To assess the significance of the observed G statistic, we computed the Z score as the difference between the observed and mean simulated values, divided by the standard deviation of the simulated values. Supplementary Data 2 lists each gene and the information used to calculate its G score. Table 1 shows the 15 genes with the highest G scores.

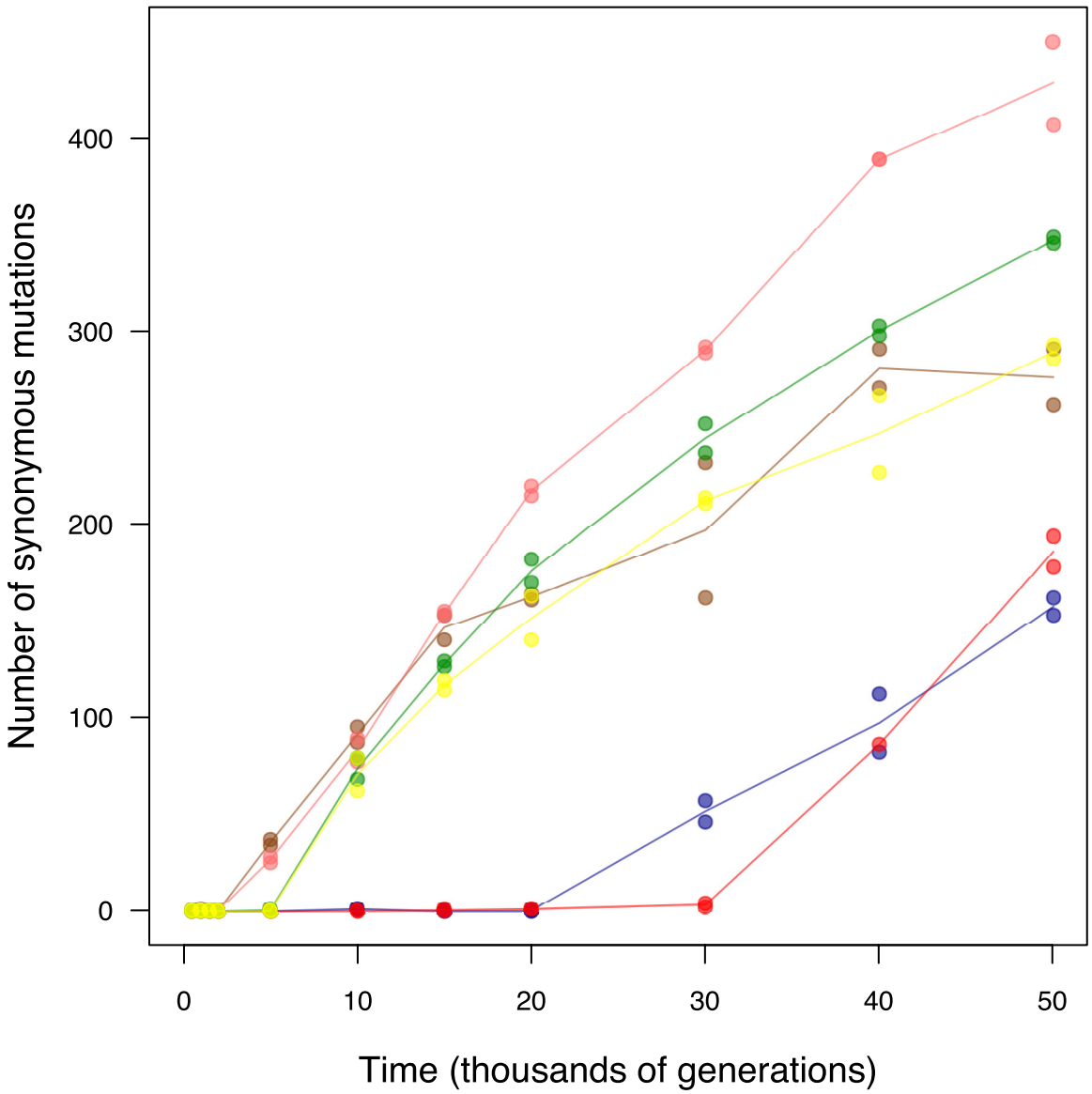
Supplementary Data 2 also shows other categories of mutation in or near each protein-coding gene including synonymous mutations, intergenic point mutations (between any particular gene and one of its immediately adjacent genes), IS insertions, small indels (≤ 50 bp), large deletions (> 50 bp) and long duplications (> 50 bp). Table 2 shows the 16 genes that had the most total deletions, duplications, insertions and intergenic point mutations (that is, all mutations except synonymous and nonsynonymous mutations in the coding gene itself).

Parallel evolution in mutator lineages. We examined parallel changes in lineages that evolved point-mutation hypermutability by analysing nonsynonymous substitutions as above. To identify mutations that occurred after a lineage became hypermutable (Supplementary Data 3), we subtracted the mutations that occurred on non-mutator branches from the total mutations. This approach may result in a few mutations that arose before hypermutability being included in the counts for mutator lineages, but given the large increases in the point-mutation rate in the mutators (Fig. 1) it provides a reasonable approximation.

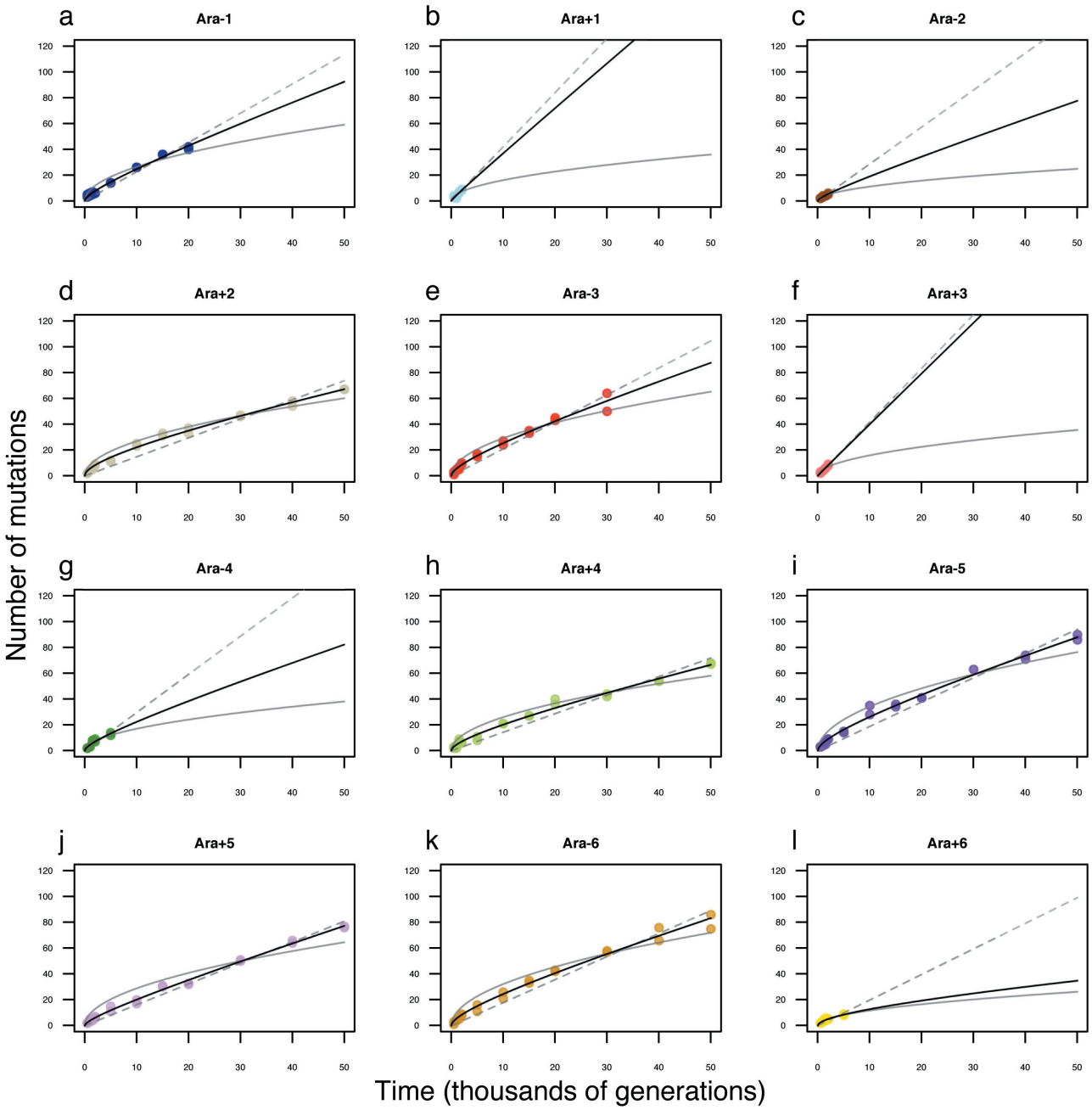
51. Daegelen, P., Studier, F. W., Lenski, R. E., Cure, S. & Kim, J. F. Tracing ancestors and relatives of *Escherichia coli* B, and the derivation of B strains REL606 and BL21(DE3). *J. Mol. Biol.* **394,** 634–643 (2009).
52. Jeong, H. *et al.* Genome sequences of *Escherichia coli* B strains REL606 and BL21(DE3). *J. Mol. Biol.* **394,** 644–652 (2009).
53. Luria, S. E. & Delbrück, M. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28,** 491–511 (1943).
54. Blount, Z. D., Borland, C. Z. & Lenski, R. E. Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **105,** 7899–7906 (2008).
55. Wielgoss, S. *et al.* Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with *Escherichia coli*. *G3 (Bethesda)* **1,** 183–186 (2011).
56. Raeside, C. *et al.* Large chromosomal rearrangements during a long-term evolution experiment with *Escherichia coli*. *mBio* **5,** e01377–14 (2014).
57. Maddamsetti, R. *et al.* Synonymous genetic variation in natural isolates of *Escherichia coli* does not predict where synonymous substitutions occur in a long-term experiment. *Mol. Biol. Evol.* **32,** 2897–2904 (2015).
58. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5,** R12 (2004).
59. Desper, R. & Gascuel, O. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J. Comput. Biol.* **9,** 687–705 (2002).
60. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20,** 289–290 (2004).



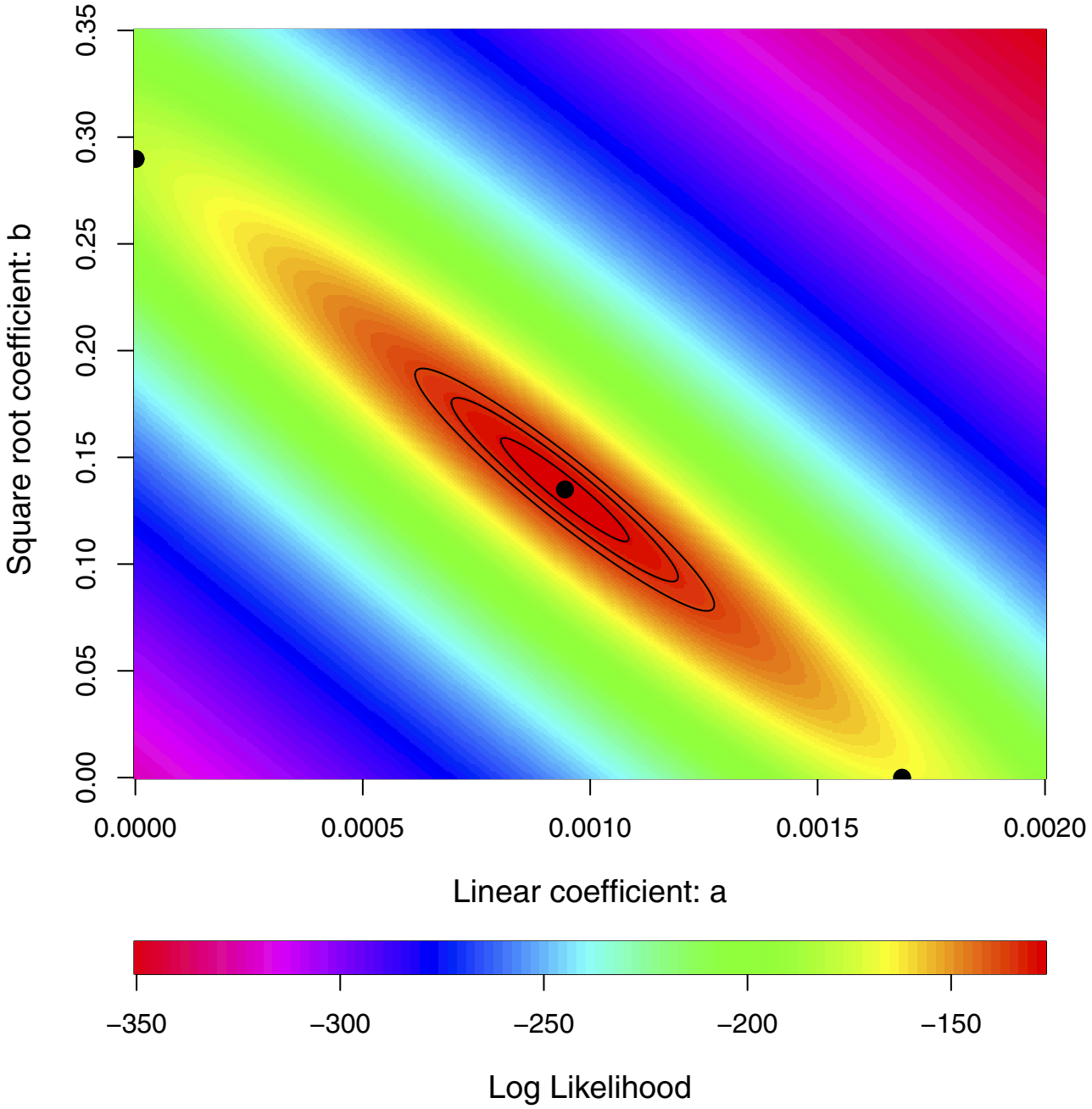
Extended Data Figure 1 | Changes in genome size during the LTEE. Box-and-whiskers plot showing the distribution of average genome length (Mb) for each of the 12 LTEE populations based on the two clones sequenced at each time point shown from 500 to 50,000 generations. The red line shows the length of the ancestral genome. The boxes are the interquartile range (IQR), which spans the second and third quartiles of the data (25th to 75th percentiles); the thick black lines are medians; the whiskers extend to the outermost values that are within 1.5 times the IQR; and the points show all outlier values beyond the whiskers.



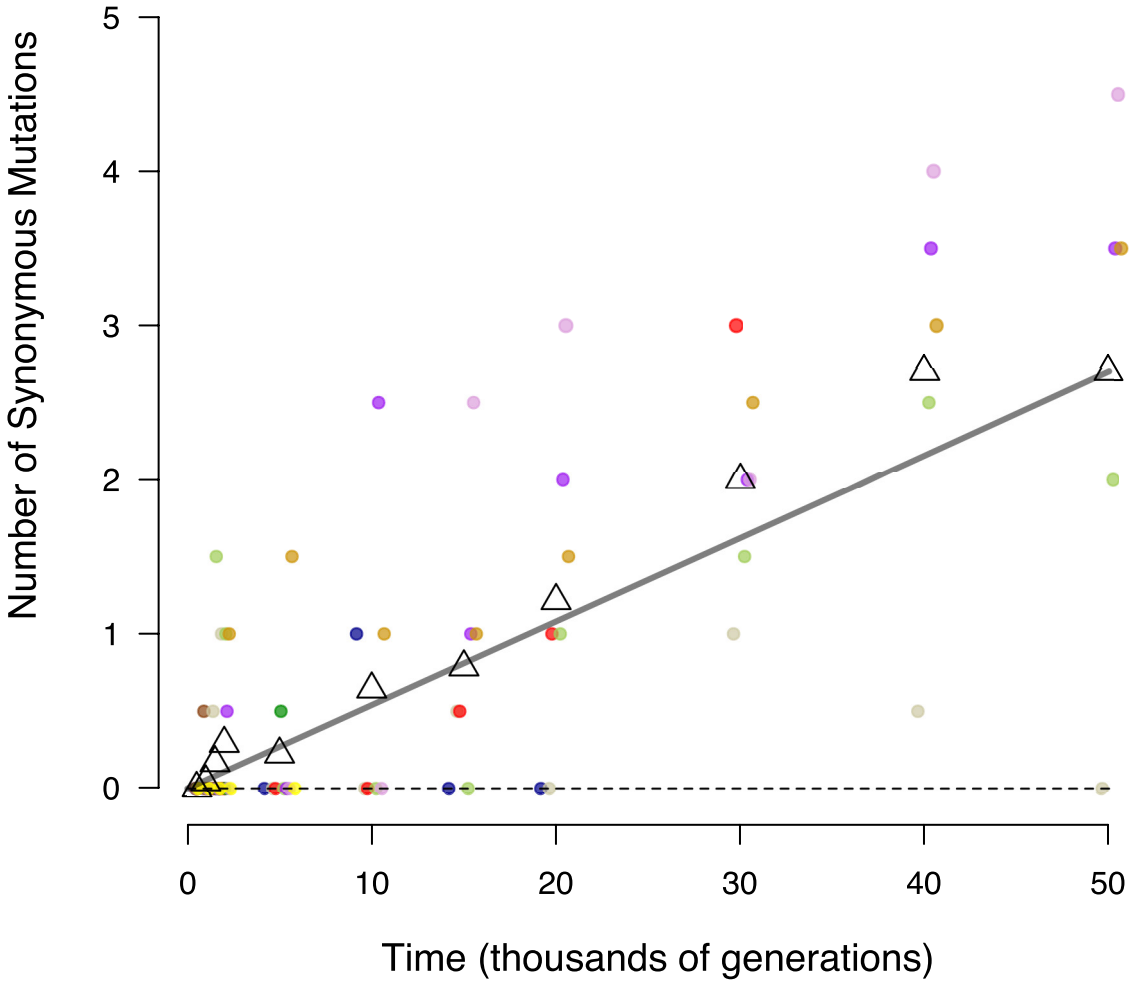
Extended Data Figure 2 | Accumulation of synonymous mutations in populations that evolved point-mutation hypermutability. Each symbol shows a sequenced genome from a hypermutable lineage. Colours are the same as those in Fig. 1. The accumulation of synonymous substitutions serves as a proxy for the underlying point-mutation rate. All four of the populations that became hypermutable before 10,000 generations accumulated synonymous mutations at higher rates between 10,000 and 20,000 generations than between 40,000 and 50,000 generations, indicating the evolution of reduced mutability.



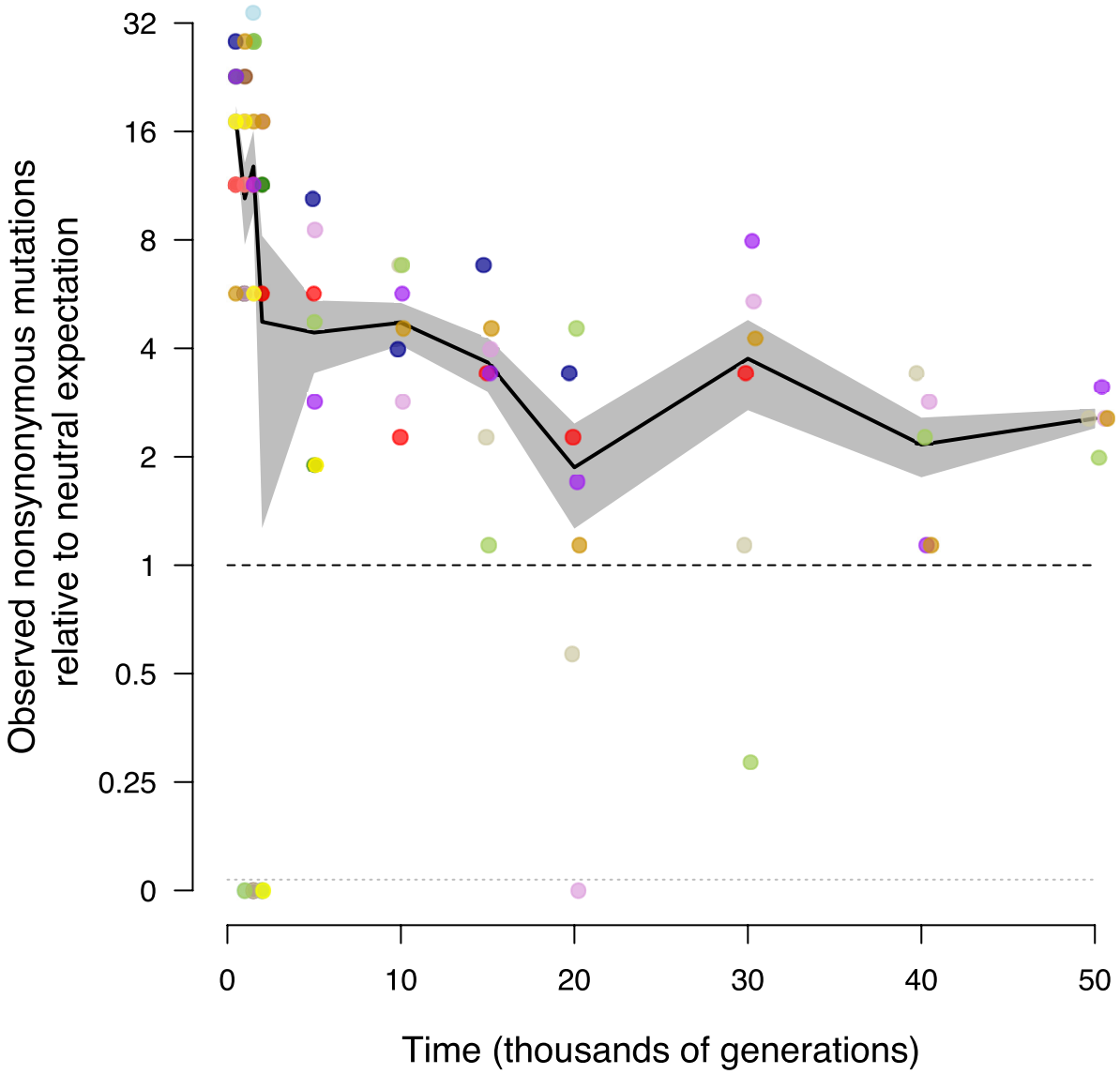
Extended Data Figure 3 | Alternative models fit to trajectory of genome evolution for each LTEE population. **a**, Ara−1. **b**, Ara+1. **c**, Ara−2. **d**, Ara+2. **e**, Ara−3. **f**, Ara+3. **g**, Ara−4. **h**, Ara+4. **i**, Ara−5. **j**, Ara+5. **k**, Ara−6. **l**, Ara+6. Each symbol shows the total mutations in a sequenced genome; in many cases, the symbols for the two genomes from the same population and generation are not distinguishable because they have the same, or almost the same, number of mutations. For the populations that evolved hypermutability, data are shown only for time points before mutators arose. In each panel, the dashed grey line shows the best fit to the linear model; the solid grey curve shows the best fit to the square-root model; and the solid black curve shows the best fit to the composite model with both linear and square-root terms.



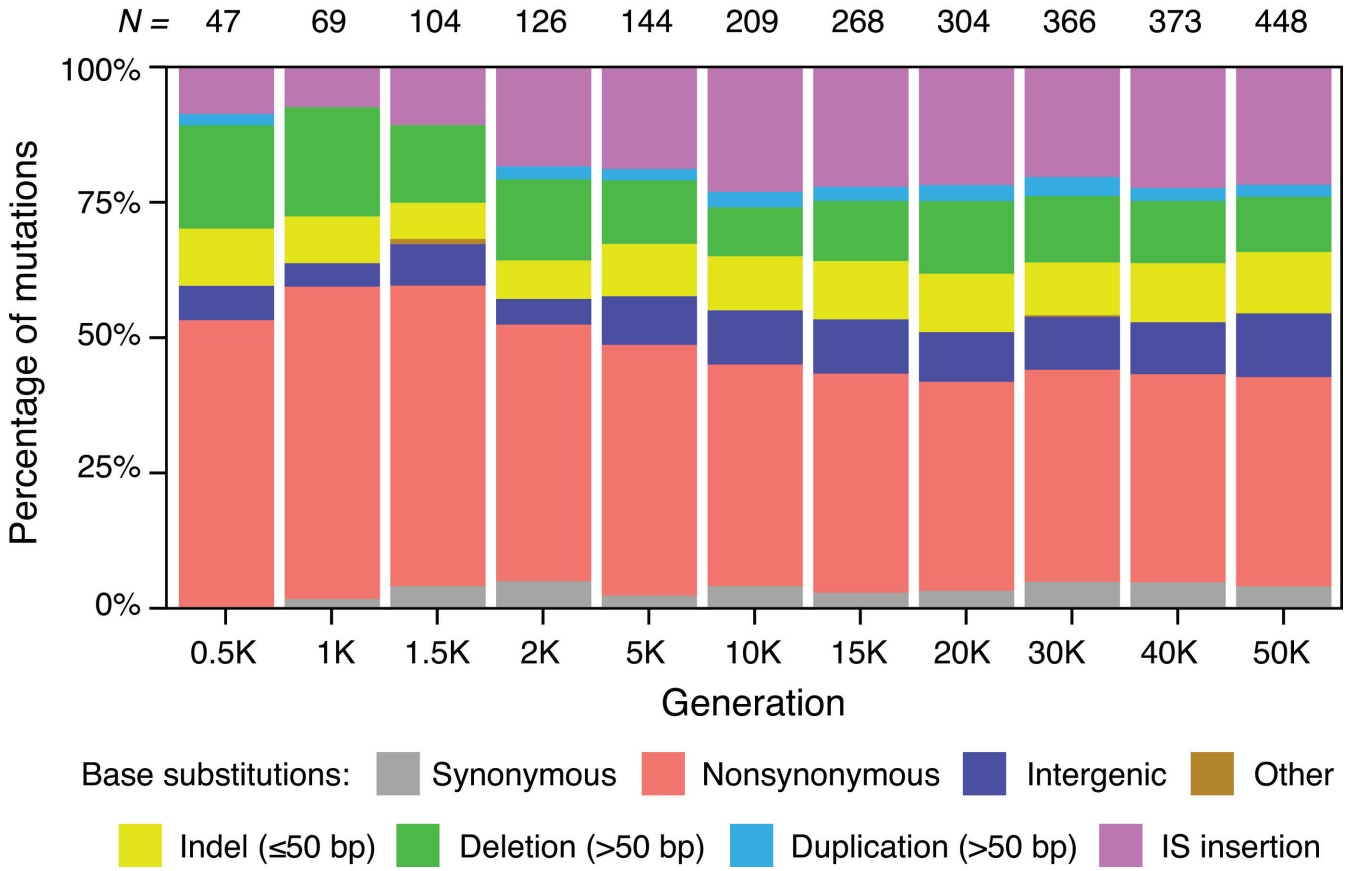
Extended Data Figure 4 | Uncertainty in parameter estimation for the model describing the rates of accumulation for neutral and beneficial mutations. Contours show relative likelihoods for simultaneously estimating the linear and square-root coefficients from the observed numbers of mutations that accumulated over time in non-mutator and premutator lineages (Fig. 3). The black central point shows the maximum likelihood estimates, and the three black contours show solutions 2, 6 and 10 log units away. The points on the horizontal and vertical axes show values for the best one-parameter models.



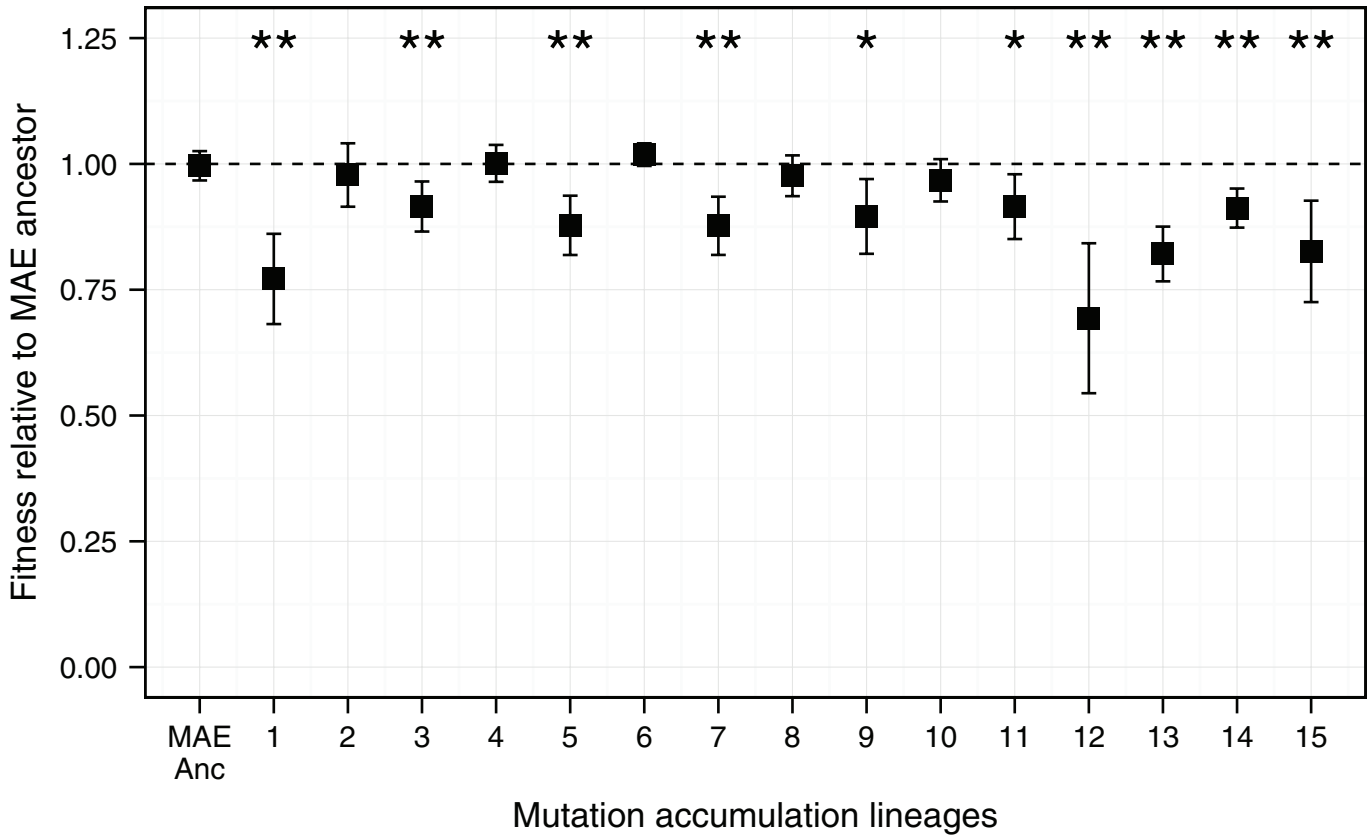
Extended Data Figure 5 | Accumulation of synonymous substitutions in non-mutator lineages. Each filled symbol shows the mean number of synonymous mutations in the (usually two) non-mutator genomes from an LTEE population that were sequenced at that time point; non-integer values can occur if the two genomes have different numbers. Small horizontal offsets were added so that overlapping points are visible. Colours are the same as in Fig. 1. Open triangles show the grand means of the replicate populations. The grey line extends from the intercept to the final grand mean. The slope of that line was used to scale the relative rates of synonymous, nonsynonymous and intergenic point mutations in Fig. 4.



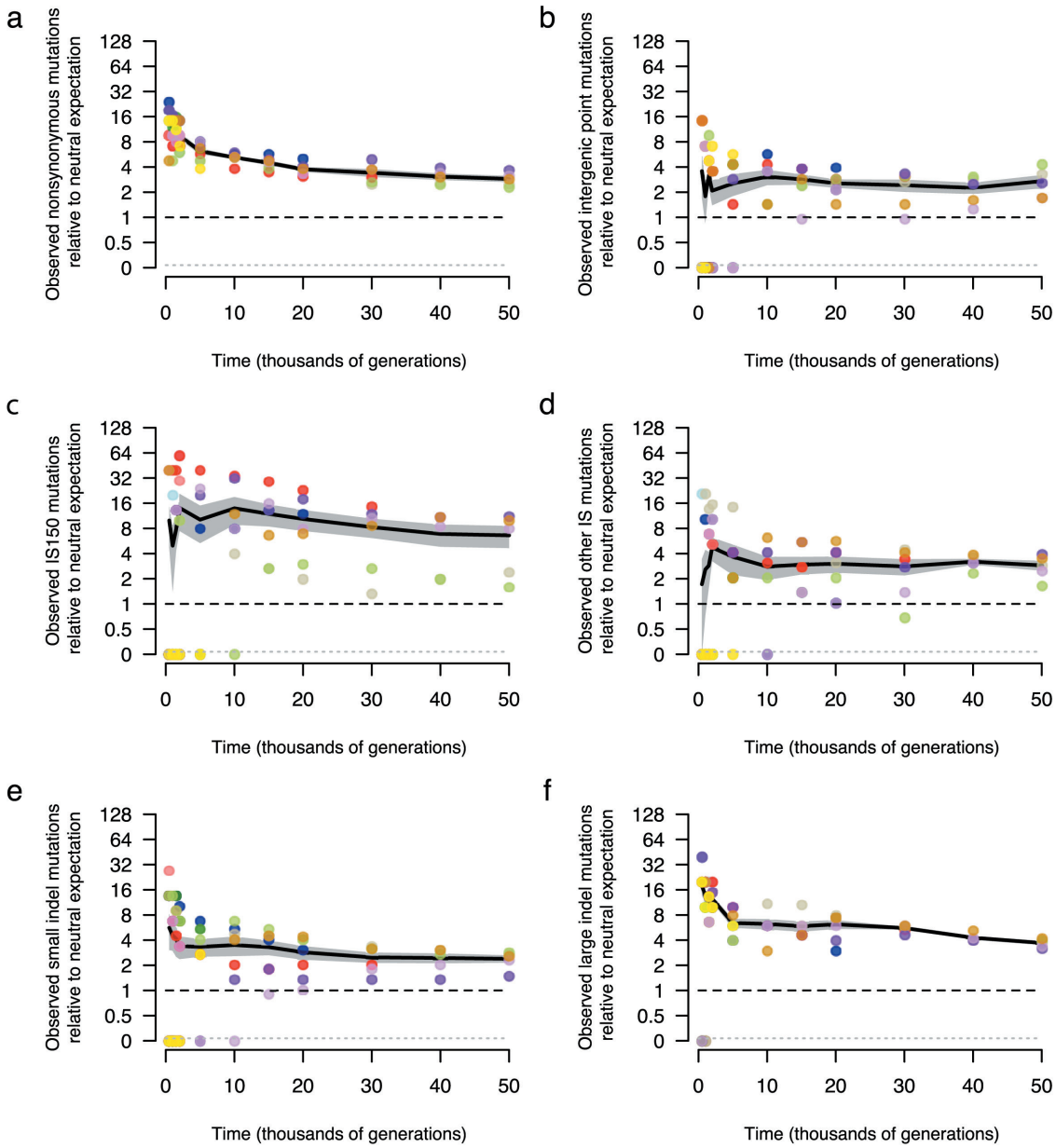
Extended Data Figure 6 | Temporal trend in accumulation of nonsynonymous mutations relative to the neutral expectation in non-mutator lineages. Interval-specific accumulation of nonsynonymous mutations calculated from changes in the total number of nonsynonymous mutations between successive samples. As with the cumulative data in Fig. 4b, values are scaled by the average rate of accumulation for synonymous mutations over 50,000 generations, after adjusting for the numbers of genomic sites at risk for nonsynonymous and synonymous mutations. Each point shows the average rate calculated for a non-mutator or premutator population; small horizontal offsets were added so that overlapping points are visible. Note the discontinuous scale; populations with no additional mutations over an interval are plotted below. Colours are the same as in Fig. 1. Black lines connect grand means; the grey shading shows standard errors calculated from the replicate populations.



Extended Data Figure 7 | Mutational spectrum for non-mutator lineages in the LTEE. Shaded bars show the distribution of different types of genetic change for all independent mutations found in the set of non-mutator clones that were sequenced at each generation. The total number of mutations in this set at each time point (*N*) is shown above each column. Base substitutions are divided into synonymous, nonsynonymous, intergenic, and other categories; the nonsynonymous category includes nonsense mutations, and the 'other' category includes rare point mutations in noncoding RNA genes and pseudogenes.



Extended Data Figure 8 | Changes in fitness of MAE lines after 550 single-cell bottlenecks and ~13,750 generations. Each point shows the mean fitness based on nine competition assays between the MAE ancestor (REL1207) or one of the 15 MAE lineages (JEB807–JEB821) and the Ara$^-$ variant of the MAE ancestor (REL1206). One-day competition assays were performed using the standard procedures and same conditions as for the LTEE[16,17]. Error bars show 95% confidence intervals. *$P < 0.05$, **$P < 0.01$, based on two-tailed t-tests of the null hypothesis that relative fitness equals 1. Ten of the fifteen MAE lines experienced significant fitness declines, while none had significant gains.



Extended Data Figure 9 | Trajectories for mutations by class in the LTEE in comparison with neutral expectations based on the MAE. **a**–**f**, Accumulation of nonsynonymous mutations (**a**), intergenic point mutations (**b**), IS*150* insertions (**c**), all other IS-element insertions (**d**), small indels (**e**) and large indels (**f**). Colours are the same as in Fig. 1. All values are expressed relative to the rate at which synonymous mutations accumulated in non-mutator LTEE lineages over 50,000 generations (Fig. 4a), and then scaled by the ratio of the number of the indicated class of mutation relative to the number of synonymous mutations in the MAE lines. In all panels, each symbol shows a non-mutator or premutator population. Note the discontinuous scale, in which populations with no mutations of the indicated type are plotted below. Black lines connect grand means over the replicate LTEE populations; the grey shading shows the corresponding standard errors.

ARTICLE

doi:10.1038/nature18933

A multi-modal parcellation of human cerebral cortex

Matthew F. Glasser[1], Timothy S. Coalson[1]*, Emma C. Robinson[2,3]*, Carl D. Hacker[4]*, John Harwell[1], Essa Yacoub[5], Kamil Ugurbil[5], Jesper Andersson[2], Christian F. Beckmann[6,7], Mark Jenkinson[2], Stephen M. Smith[2] & David C. Van Essen[1]

Understanding the amazingly complex human cerebral cortex requires a map (or parcellation) of its major subdivisions, known as cortical areas. Making an accurate areal map has been a century-old objective in neuroscience. Using multi-modal magnetic resonance images from the Human Connectome Project (HCP) and an objective semi-automated neuroanatomical approach, we delineated 180 areas per hemisphere bounded by sharp changes in cortical architecture, function, connectivity, and/or topography in a precisely aligned group average of 210 healthy young adults. We characterized 97 new areas and 83 areas previously reported using post-mortem microscopy or other specialized study-specific approaches. To enable automated delineation and identification of these areas in new HCP subjects and in future studies, we trained a machine-learning classifier to recognize the multi-modal 'fingerprint' of each cortical area. This classifier detected the presence of 96.6% of the cortical areas in new subjects, replicated the group parcellation, and could correctly locate areas in individuals with atypical parcellations. The freely available parcellation and classifier will enable substantially improved neuroanatomical precision for studies of the structural and functional organization of human cerebral cortex and its variation across individuals and in development, aging, and disease.

Neuroscientists have long sought to subdivide the human brain into a mosaic of anatomically and functionally distinct, spatially contiguous areas (cortical areas and subcortical nuclei), as a prerequisite for understanding how the brain works. Areas differ from their neighbours in microstructural architecture, functional specialization, connectivity with other areas, and/or orderly intra-area topographic organization (for example, the map of visual space in visual cortical areas)[1–3]. Accurate parcellation provides a map of where we are in the brain, enabling efficient comparison of results across studies and communication among investigators; as a foundation for illuminating the functional and structural organization of the brain; and as a means to reduce data complexity while improving statistical sensitivity and power for many neuroimaging studies.

The human cerebral cortex has been estimated to contain anywhere from ~50 (ref. 1) to ~200 (refs 3, 4) areas per hemisphere. However, attaining a consensus whole-cortex parcellation has been difficult because of practical and technical challenges that we address here.

Most previous parcellations were based on only one neurobiological property (such as architecture, function, connectivity or topography), and many cover only part of the cortex. Combining multiple properties provides complementary as well as confirmatory information, as different properties distinguish different sets of areal boundaries, and more confidence can be placed in boundaries that are consistent across multiple independent properties. We analysed all four properties across all of neocortex in both hemispheres, using new or refined methods applied to the uniquely rich repository of exceptionally high-quality magnetic resonance imaging (MRI) data provided by the Human Connectome Project (HCP), which benefited from major advances in image acquisition and preprocessing[5–8]. Architectural measures of relative cortical myelin content and cortical thickness were derived from T1-weighted (T1w) and T2-weighted (T2w) structural images[5,9,10]. Cortical function was measured using task functional MRI (tfMRI) contrasts from seven tasks[11]. Resting-state functional MRI (rfMRI) revealed functional connectivity of entire cortical areas plus topographic organization within some areas.

Previous parcellations typically used either fully automated algorithmic approaches, or else manual or partly automated neuroanatomical approaches in which neuroanatomists delineated areal borders, documented areal properties, and identified areas after consulting prior literature. Here we combined both approaches. For the initial parcellation, we adapted a successful observer-independent semi-automated neuroanatomical approach for generating post-mortem architectonic parcellations[12,13] to non-invasive neuroimaging. We used an algorithm to delineate potential areal borders (transitions in two or more of the cortical properties described above), which two neuroanatomists (authors M.F.G. and D.C.V.E.) then interpreted, documenting areal properties and identifying areas relative to the extant neuroanatomical literature. We then used a fully automated algorithmic approach, training a machine-learning classifier to delineate and identify cortical areas in individual subjects based on multi-modal areal fingerprints, allowing the parcellation to be replicated in new subjects and studies.

Prior parcellations have either used small numbers of individuals or group averages that are 'blurry' from inaccurate alignment of brain areas across subjects. We aligned cortical data using 'areal features', including maps of relative myelin content and resting state networks that are more closely tied to cortical areas than are the folding patterns typically used for alignment[14]. The markedly improved intersubject cortical alignment using cortical folding, myelin, and resting state fMRI enabled us to generate the 'typical subject's' parcellation from a highly detailed 210-subject group average data set.

[1]Department of Neuroscience, Washington University Medical School, Saint Louis, Missouri 63110, USA. [2]FMRIB Centre, Nuffield Department of Clinical Neurosciences, John Radcliffe Hospital, University of Oxford, Oxford OX3 9DU, UK. [3]Department of Computing, Imperial College, London SW7 2AZ, UK. [4]Department of Biomedical Engineering, Washington University, Saint Louis, Missouri 63110, USA. [5]Center for Magnetic Resonance Research (CMRR), University of Minnesota, Minneapolis, Minnesota 55455, USA. [6]Donders Institute for Brain, Cognition and Behavior, Radboud University, Nijmegen 6525 EN, The Netherlands. [7]Department of Cognitive Neuroscience, Radboud University Medical Centre Nijmegen, Postbus 9101, Nijmegen 6500 HB, The Netherlands.
*These authors contributed equally to this work.

Group map reproducibility of fine details

We analysed two independent groups of HCP subjects—210P ('parcellation') and 210V ('validation')—aligned using areal-feature-based registration (called 'MSMAll', see Methods section on image preprocessing). Figure 1 illustrates the consistency of fine spatial patterns for maps reflecting relative myelin content (left panels) and a task-fMRI language-related activation (right panels). The maps are strikingly similar across the 210P and 210V groups, including variations in relative myelin content within the primary somatosensory cortex related to somatotopic organization (white and black arrows in left panels, see legend) and small features in the task fMRI maps (white ellipses in right panels). Supplementary Results and Discussion 1.1 and associated Supplementary Figs 1–5 include more examples of such cross-group consistency for architecture (myelin and cortical thickness), function (tfMRI contrast maps), and two resting state connectivity measures.

Because of the areal-feature-based alignment, maps of average cortical folding in this study are much blurrier than are maps of areal properties because folding patterns and area locations are imperfectly correlated[4,12,14] (for example, compare Supplementary Fig. 7e and 7j (group average and individual subject folding) in Supplementary Results and Discussion 1.3). Group average folding patterns remain sharp mainly in early sensory areas, where areal locations and folds are tightly correlated (for example, central and calcarine sulci, see Supplementary Fig. 1, rows 3 and 4, in Supplementary Results and Discussion 1.1). The regional difference in sharpness between maps of areal properties and folding highlights the importance of alignment based on areal features, rather than folding patterns, as a prerequisite for accurately parcellating group average data. The high spatial resolution of the HCP's MRI images and lack of aggressive spatial smoothing[10] prior to group averaging also contribute to making our maps substantially sharper than those from traditional neuroimaging studies.

Quantitatively, the 210P and 210V group average datasets were highly correlated across the cortical surface ($r = 0.998$ for myelin; $r = 0.994$ for cortical thickness after correction for folding-related effects; $r = 0.996$ and $r = 0.979$ for two folding-related measures (FreeSurfer's 'sulc' and 'curv', respectively); $r = 0.995$, $r = 0.984$ and $r = 0.944$ for the maximum, median and minimum, respectively, of the task fMRI contrasts, and a median reproducibility of $r = 0.989$ for two measures of resting state connectivity). These excellent map reproducibilities provide confidence that the parcellation will reflect the areal pattern of typical subjects in the healthy young adult population. See Methods section on modalities for parcellation, and Supplementary Methods 1.3–3.4 for the methods used to generate these maps.

A 180-area group average parcellation

To identify transitions representing candidate areal boundaries, we designed and implemented a semi-automated, quantitative approach adapted for multi-modal neuroimaging data represented on two-dimensional cortical surface models (see Methods section on the gradient-based parcellation approach and Supplementary Methods 4.1–5.3). The approach is similar in spirit to a highly successful semi-automated observer-independent approach[13,15]. However, instead of objectively identifying potential areal borders in postmortem histological sections, we identified them algorithmically on the cortical surface by computing the first derivative of each areal feature map (its spatial gradient magnitude)[16]. Candidate borders were then interpreted by the neuroanatomists to exclude artefacts. Each area's properties were documented (in the Supplementary Neuroanatomical Results), and putative areas were related to the extant neuroanatomical literature.

These semi-automated approaches contrast with classical observer-dependent parcellation approaches[1,3] that have relied on visual inspection to locate often subtle transitions in cortical architecture and with some modern observer-dependent retinotopic parcellation methods[17,18]. They also differ from fully automated, unsupervised methods[19–21] in which the outcomes depend heavily on algorithmic input parameters (for example, thresholds or number of requested clusters) and are not validated by a neuroanatomist.

Area 55b illustrates our multi-modal gradient-based parcellation approach using gradients of three areal feature maps (see Fig. 2). Area 55b is a small, elongated, and notably distinct area (outlined in black or white) bounded by the frontal eye field (FEF) and premotor eye field (PEF), primary motor cortex (4), ventral premotor cortex (6v), and prefrontal areas 8Av and 8C. In the myelin map (Fig. 2a), area 55b is lightly myelinated and lies between moderately myelinated areas FEF (above) and PEF (below), just anterior to heavily myelinated primary motor cortex (area 4). Thus, area 55b is surrounded on three sides by myelin gradients (Fig. 2e). Area 55b is strongly activated in the 'Story versus Baseline' task contrast from the HCP's 'LANGUAGE' task (Fig. 2b) and is entirely surrounded by a strong gradient for this task contrast (Fig. 2f). It also has distinctive functional connectivity, as revealed by a seed location (lightly myelinated area PSL) selectively connected with 55b (Fig. 2c) and a different seed location (heavily myelinated area LIPv) strongly connected with FEF and PEF (Fig. 2d) but not with 55b. The result is strong mean gradients in dense functional connectivity surrounding 55b (Fig. 2g). Ref. 22 illustrated area 55b on a schematic surface map (Fig. 2h) as a lightly myelinated area bounded on three sides by more heavily myelinated areas. Because of the similarity to the dorsal portion of 55b in ref. 22, we use the same name.

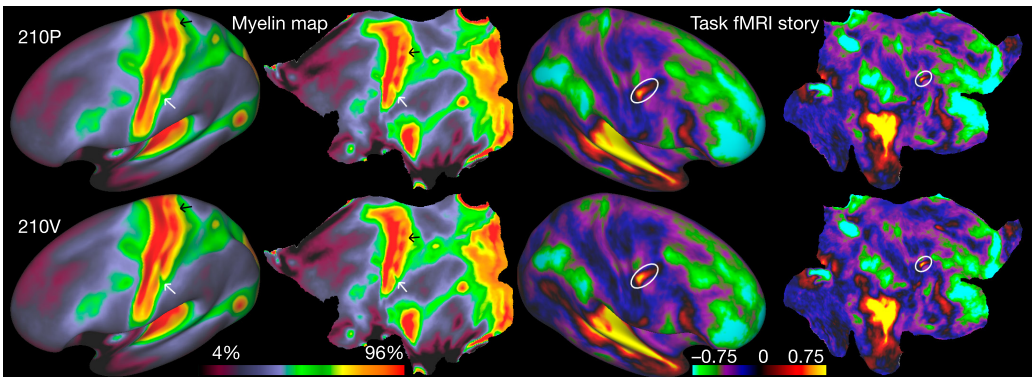


Figure 1 | Consistency of fine spatial details in independent group averages. Relative myelin content maps (left hemisphere) and task fMRI contrast beta maps from the LANGUAGE story contrast (right hemisphere) on inflated (columns 1 and 3) and flattened surfaces (columns 2 and 4). Rows 1 and 2 are the group averages of the 210P and 210V data sets, respectively. White and black arrows indicate consistent variations in myelin content within primary somatosensory cortex that are correlated with somatotopy (see Supplementary Neuroanatomical Results 6 and Supplementary Neuroanatomical Results Fig. 8). The white oval indicates a small, sharp, and reproducible feature in the right hemisphere of the LANGUAGE story contrast. Relative myelin content will hereafter be referred to as myelin (see legend of Supplementary Fig. 1 in Supplementary Results and Discussion 1.1). Data at http://balsa.wustl.edu/WDpX.

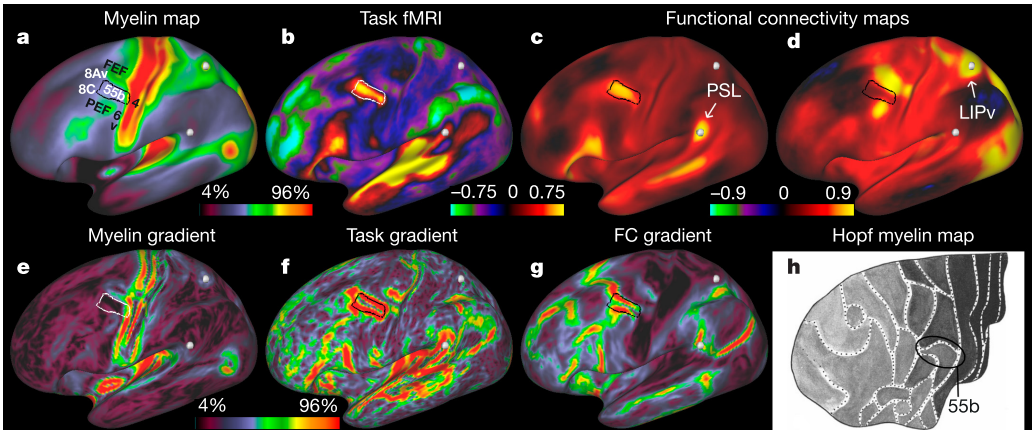


Figure 2 | Parcellation of exemplar area 55b using multi-modal information. The border of 55b is indicated by a white or black outline. **a**, Myelin map. **b**, Group average beta map from the LANGUAGE Story versus Baseline task contrast. **c**, **d**, Functional connectivity correlation maps from a seed in area PSL (white sphere, arrow) (**c**) and a seed in area LIPv (white sphere, arrow) (**d**). **e**, Gradient magnitude of the myelin map shown in **a**. **f**, Gradient magnitude of the LANGUAGE Story versus Baseline task contrast shown in **b**. **g**, Mean gradient magnitude of the functional connectivity dense connectome (see section on modalities for parcellation in the Methods). **h,** A dorsal schematic view of the prefrontal cortex as parcellated in ref. 22, in which shading indicates the amount of myelin found using histological stains of cortical grey matter. Data at http://balsa.wustl.edu/Qv4P.

To generate the complete parcellation of 180 areas and area complexes in each hemisphere, we adopted a systematic, objective, and quantitative approach (see the gradient-based parcellation approach section in the Methods and in Supplementary Methods 5.1–5.3). Our major criteria, met in nearly all cases, included: (i) spatially overlapping gradient 'ridges' between each pair of areas for at least two independent areal feature maps; (ii) similar gradient ridges present in roughly corresponding locations in both hemispheres; (iii) gradients that were not correlated with artefacts; and (iv) robust and statistically significant cross-border differences in the feature maps. Another consideration (but not a requirement) was whether published evidence exists for a boundary in an approximately corresponding location. Studies with publicly available parcellations registered onto atlas surfaces[4] were directly compared with our data; however, most regions required indirect comparisons with published figures (for example, Fig. 1h). Initial areal boundaries meeting these criteria were delineated by two neuroanatomists (authors M.F.G. and D.C.V.E.).

In a second computational stage, the path of each manually drawn border was optimized algorithmically using gradients of the most informative feature maps selected by the neuroanatomists (those with visually obvious gradients and differences across the border). These feature maps were confirmed to have robust and statistically significant differences across the final border. The semi-automated gradient-based parcellation approach is further described in Supplementary Methods 5.1–5.3), and the entire semi-automated process is illustrated for area V1 in Supplementary Neuroanatomical Results 1; other sections of this document describe and illustrate the information used to delineate and the literature used to name all 180 cortical areas.

Figure 3 shows the multi-modal cortical parcellation in the left and right hemispheres on inflated and flattened surfaces, with areal

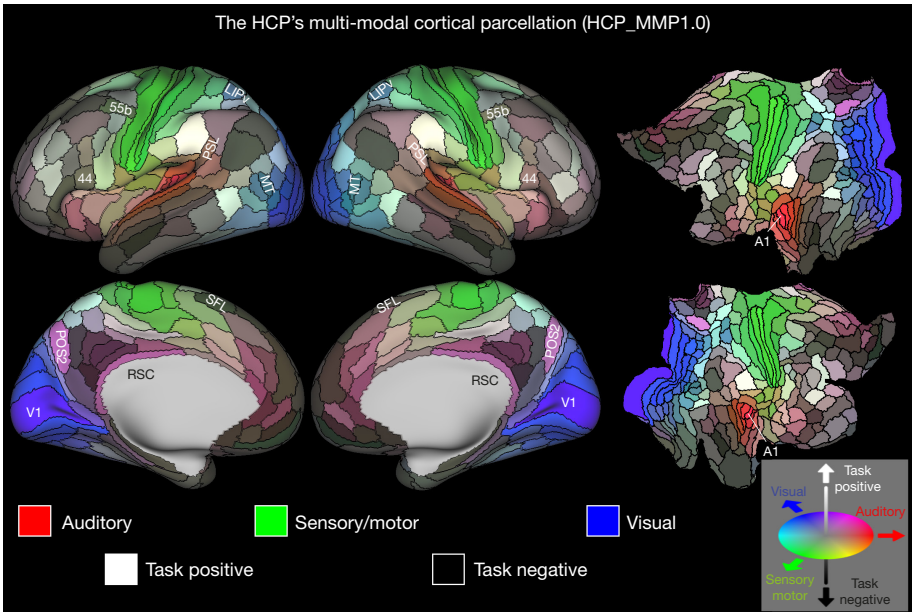


Figure 3 | The HCP's multi-modal parcellation, version 1.0 (HCP_MMP1.0). The 180 areas delineated and identified in both left and right hemispheres are displayed on inflated and flattened cortical surfaces. Black outlines indicate areal borders. Colours indicate the extent to which the areas are associated in the resting state with auditory (red), somatosensory (green), visual (blue), task positive (towards white), or task negative (towards black) groups of areas (see Supplementary Methods 5.4). The legend on the bottom right illustrates the 3D colour space used in the figure. Data at http://balsa.wustl.edu/WN56.

boundaries delineated by black contours. A total of 180 areas and areal complexes per hemisphere is near the higher end of earlier estimates noted above[3,4]. We consider 180 likely to be a lower bound, as some parcels are probably complexes of multiple areas (for example, based on finer-grained published parcellations, and other regions that suffer from reduced sensitivity due to fMRI signal loss). Many areas (83/180) were assigned names based on published parcellations from dozens of separate studies that used a variety of invasive or specialized methods (see Table 2 in Supplementary Neuroanatomical Results), reflecting how far the field has been from a consensus neuroanatomical parcellation. Some of the newly described 97 areas have de novo names (for example, DVT for the dorsal visual transitional area), while others represent finer-grained parcellations of previously reported areas (for example, area 31 into areas 31a, 31pd, and 31pv). A few represent complexes in which a published finer grained parcellation was not visible in our data (for example, areas 29 and 30 combined into area the retrosplenial complex (RSC)), but these may be again subdivided once higher resolution data is available. The 180 areas differ widely in their shapes, sizes, and the positions of their borders relative to cortical folds.

The parcellation in Fig. 3 is coloured to reflect each area's degree of association in the resting state (determined using multiple regression, see Supplementary Methods 5.4) with five functionally specialized groups of areas: early auditory (red), early somatosensory/motor (green), and early visual areas (blue). These represent the three dominant input streams to the brain. Also used were two core groups of cognitive areas that are strongly anti-correlated in our data, the task positive network (towards white) and task negative (also called the default mode) network (towards black). Hence, the strongly bluish, greenish, and reddish regions are predominantly but not exclusively associated with visual, somatosensory-motor, and auditory processing, respectively. Qualitatively, the predominantly unimodal regions appear to collectively occupy less than half of the neocortical sheet. Areas probably more strongly biomodal include blue-green areas such as LIPv and MT (visual and somatosensory-motor) and purple areas such as POS2 and RSC (visual and auditory). The remaining regions form a complex mosaic, with some intermixing of lighter (task-positive) and darker (task-negative) areas along with many lighter or darker pastel hues suggestive of 'cognitive' areas that may be preferentially associated with one or another sensory modality. The bilateral symmetry of functional organization is striking, in that nearly all areas have qualitatively very similar hues in the left and right hemispheres. However, interesting colour asymmetries occur in a few areas, especially language-related areas 55b, PSL, SFL, and 44 and their right hemisphere homologues, which also have asymmetric task-fMRI functional profiles (see Supplementary Neuroanatomical Results 8, 15, 21 and 22).

Internal heterogeneity is evident in some cortical areas, particularly those with topographically organized representations. In the somatosensory-motor strip (largely architecturally defined somatosensory and motor areas 3a, 3b, 1, 2, and 4), we identified five clearly defined topographic subareas in resting state and task fMRI data (see Supplementary Neuroanatomical Results 6 and the associated Supplementary Fig. 8). In this parcellation we treat topographic subdivisions as 'subareas' rather than calling them full 'areas'. For visual cortex, its visuotopic organization revealed a set of hemifield representations in each hemisphere, something not achieved in previous unsupervised resting state functional connectivity-based parcellations[19–21,23]. Also, ultrahigh-field MRI reveals sub-areal cortical organization along both laminar[24,25], and columnar[26,27] axes, so our parcellation represents one of many important levels of granularity in brain organization.

Cross-validation of the parcellation

The initial statistical analysis used in the semi-automated parcellation was circular, to the extent that the 210P dataset was used for both creating and testing the parcellation. Hence, we carried out an additional statistical cross-validation using the 210V dataset and a comprehensive set of feature maps (see the statistical cross validation of the multi-modal parcellation section of the Methods and Supplementary Results and Discussion 1.2). This analysis also reveals which areal properties were most useful in defining areal boundaries (a condensed representation of the detailed information provided in the Supplementary Neuroanatomical Results). Supplementary Fig. 6 in Supplementary Results and Discussion 1.2 shows four independent categories of features: cortical thickness, myelin maps, task fMRI, and resting state fMRI, and how many of these categories showed robust and statistically significant differences across each areal border. Fully 96% of areal borders had robust effect sizes (Cohen's $d > 1$) in two or more feature categories and all were statistically significant after correcting for multiple comparisons in two or more feature categories in cross-border, across-subject t-tests. Resting state fMRI was the most useful category, followed by task fMRI, myelin maps, and lastly cortical thickness, which was consistent with the neuroanatomists' observations and documentation in the Supplementary Neuroanatomical Results.

Exemplar parcellation-based analyses

Spatial smoothing is often used to increase the signal-to-noise ratio (SNR) in neuroimaging analyses, to try to compensate for inaccurate registration of brain areas, and/or to satisfy statistical assumptions. However, smoothing blurs data across boundaries between areas (on the surface) and tissue compartments (in the volume). An areal parcellation enables area-wise analyses (averaging data within each area), thereby improving SNR and statistical power without the deleterious effects of spatial smoothing (to the extent that properties within an area are uniform). Parcellation dramatically reduces data dimensionality, illustrated here using the HCP's myelin, thickness, task, and resting state data (Fig. 4).

The 'dense' (vertex-wise) myelin map shown in Fig. 4a has ~30,000 surface grey matter vertices per hemisphere, whereas a 'parcellated' myelin map (Fig. 4f) shows the same overall pattern with 180 cortical areas (vertices within an area have the same value, see also Fig. 4g for parcellated cortical thickness). Example dense and parcellated task fMRI analysis contrast maps (Figs. 4b, c, LANGUAGE Story versus Baseline) can be represented as a single column (white) in a 180-area by 86-task-contrast matrix (Fig. 4d). Parcellated analyses hold great promise for task fMRI studies, as they improve the signal-to-noise ratio by averaging fMRI time series within parcels prior to fitting the task design, increasing Z statistics (Fig. 4e). Parcellation is effectively a neurobiologically constrained smoothing approach that also increases statistical power by efficiently consolidating otherwise non-independent statistical tests. This approach will benefit studies aimed at understanding the functional and structural organization of the brain in health or disease at an area-wise level (studies that currently summarize results using three-dimensional coordinates in a standardized stereotaxic space). Parcellated analyses also aid in the clarity and efficiency of communicating results (for example: "area 55b in the left hemisphere showed a statistically significant +1% BOLD activation in my language task").

Parcellated analyses are comparably useful when characterizing structural or functional connectivity, as previously recognized[23,28]. Preprocessing of HCP data results in fMRI data represented as 'grayordinates' (cortical grey matter surface vertices and subcortical grey matter voxels[5]). A dense connectome, containing connectivity between all pairs of 91,282 grayordinates is $\sim 3.3 \times 10^5$-fold larger than an area-wise parcellated connectome for ~500 areas (connectivity between all pairs of areas), yet the parcellated connectome captures the neurobiologically relevant variance at the areal level. Parcellated connectomes are illustrated using a seed location in area PGi (black dot) for full correlation (Fig. 4h) and partial correlation (Fig. 4i) functional connectivity brain maps together with their associated parcellated connectome matrices (Fig 4j, full correlation below and partial correlation above the diagonal). In both cases, the task

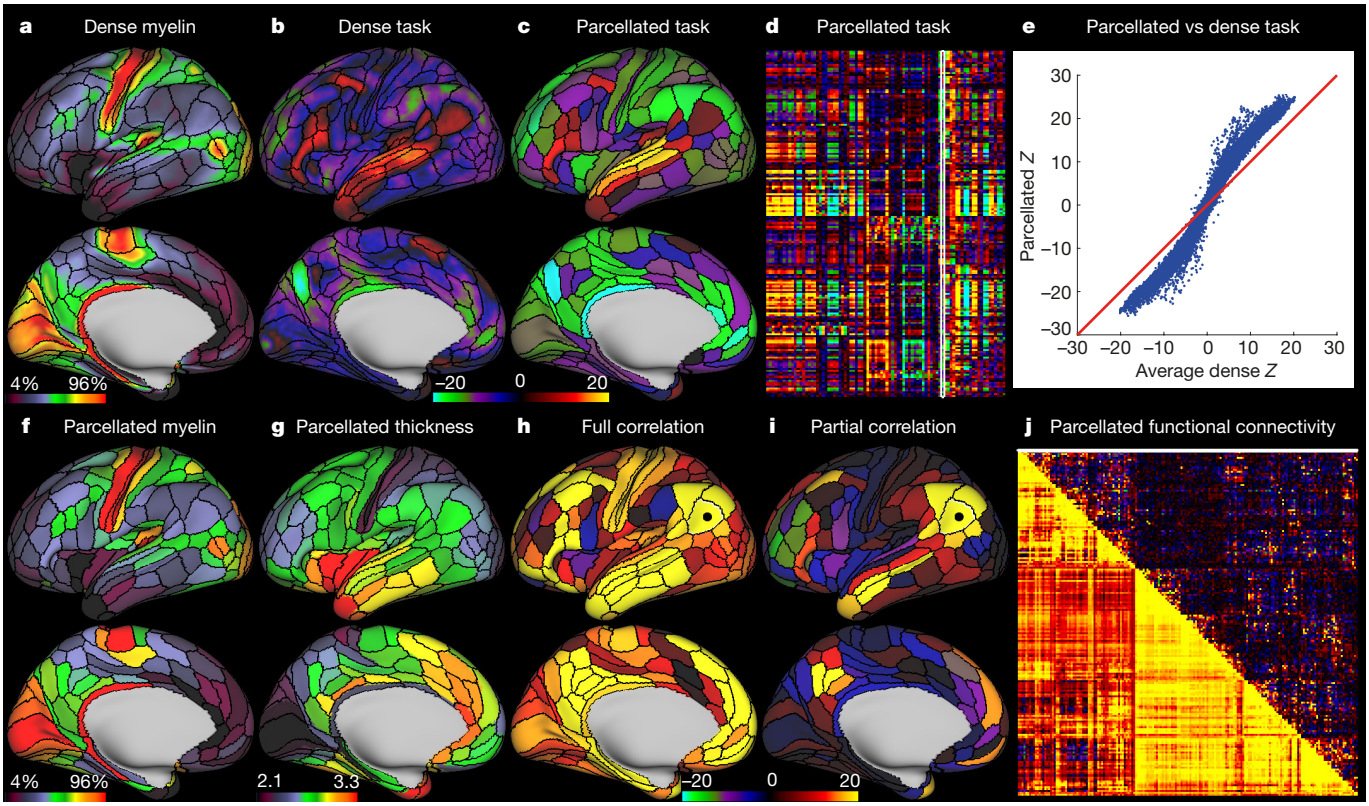


Figure 4 | Example parcellated analyses using the HCP's multi-modal cortical parcellation. a, Dense myelin maps on lateral (top) and medial (bottom) views of inflated left hemisphere. **b**, **c**, Example dense (**b**) and parcellated (**c**) task fMRI analysis (LANGUAGE story versus baseline) expressed as *Z* statistic values. **d**, The entire HCP task fMRI battery's *Z* statistics for 86 contrasts (47 unique, see section on modalities for parcellation in the Methods) analysed in parcellated form and displayed as a matrix (rows are parcels, columns are contrasts, white outline indicates the map in **c**). **e**, A major improvement in *Z* statistics from fitting task designs on parcellated time series instead of fitting them on dense time series and then parcellating afterwards (blue points are 360 parcels × 86 task contrasts; note the upward tilting deviation from the red line). **f**, Parcellated myelin maps. **g**, A parcellated folding-corrected cortical thickness map (in mm). **h**, **i**, Parcellated functional connectivity maps on the brain (seeded from area PGi, black dot). These parcellated connectomes are computed using either full or partial correlation (see Supplementary Methods 7.1). In both cases, the task negative (default mode) network is apparent. **j**, A parcellated connectome matrix view with the full correlation connectome below and the partial correlation connectome above the diagonal (white line shows the displayed partial correlation brain map). Data at http://balsa.wustl.edu/RG0x.

negative (default mode) network is evident, though the partial correlation connectome is much sparser than the full correlation connectome.

Individuals with atypical areal patterns

The precisely aligned group average multi-modal cortical parcellation represents the overall spatial arrangement of cortical areas in the 'typical' individual from a healthy young adult population. However, we found atypical topological arrangements of some areas in some individuals that are discernible across multiple modalities, including resting-state networks, task-fMRI activations, and myelin maps. Distinguishing genuinely atypical areal topologies from inadequately aligned typical patterns depended on the MSMAll areal-feature-based registration to align cortical areas precisely. We summarize key findings here and extensively characterize this important phenomenon in the Supplementary Results and Discussion 1.3.

Previously described area 55b and neighbouring areas FEF and PEF showed particularly notable individual differences in topological arrangements. For the 210P subjects, 89% showed the typical configuration in the left hemisphere (area 55b bordered by area PEF inferiorly and area FEF superiorly, as in Fig. 2), which was well aligned with group average area 55b after MSMAll registration. However, in one subgroup (4%, $n = 9$), a patch having the multi-modal characteristics of area 55b is shifted superiorly relative to the upper limb subregion of sensori-motor cortex (Supplementary Figure 7 in Supplementary Results and Discussion 1.3). In another subgroup (6%, $n = 12$), area 55b is split into two pieces by a merger of areas FEF and PEF, rather than the typical splitting of FEF and PEF by 55b (Supplementary Fig. 8 in Supplementary Results and Discussion 1.3). Such topological deviations in individual subjects' areal maps raise intriguing questions for future exploration. They also cannot be corrected by a topology-preserving registration aimed at aligning individual subjects' areas with the group average 'atlas' parcellation. Thus, we introduce an alternative fully automated cortical parcellation approach that can identify and delineate both typical and atypical areas in individual subjects that were not a part of the original 210P group.

Automated individual-subject parcellation

The semi-automated neuroanatomical approach described above is impractical for de novo individual subject parcellation of all ~1,100 HCP subjects having complete MRI datasets so as to identify the atypical areal topologies mentioned above. Instead, we developed an automated method for generating individual subject parcellations based on a supervised machine learning classifier previously used to identify resting state functional networks in individual subjects[29]. In our case, the areal classifier learns the multi-modal 'areal fingerprint' of each cortical area that distinguishes it from surrounding cortex. Based on multi-modal feature maps that represent the areal properties of architecture, function, connectivity, and topography, the areal classifier returns a prediction (0% to 100%) that each area exists at a given cortical surface vertex. The highest prediction value across areas at each vertex is used to generate the individual subject

parcellation (see the cortical areal classifier section in Methods and in Supplementary Methods 6.1–6.8). Once trained using the 210P subjects (and a separate '29T' group of test subjects, see the subjects and acquisitions section of the Methods), the areal classifier should be able to use only the multi-modal areal fingerprints that it has learned to reproduce the parcellation in an independent group of validation subjects (210V).

A critical early test of the areal classifier was whether it could accurately and reliably map areas that are not aligned with the population-based atlas parcellation after MSMAll areal-feature-based alignment (see Supplementary Results and Discussion 1.4). Examples of successful classification of areas 55b, FEF, and PEF are shown in Supplementary Fig. 9 of the Supplementary Results and Discussion for typical subjects, shifted 55b subjects, and split 55b subjects. In each illustrated case, the classifier correctly identified 55b and its neighbours (as assessed by the neuroanatomists' inspection of the multi-modal areal features shown in the figure). Supplementary Fig. 10 in the Supplementary Results and Discussion 1.4 shows that these atypical 55b topologies and classifications are stable across widely spaced repeat scanning sessions in a 'test–retest' group of 27 subjects (see Methods section on subjects and acquisition).

Areal detection and parcellation consistency

Another critical test of both the parcellation and the areal classifier is the classifier's performance in detecting the 180 cortical areas in individual subjects, particularly in independent validation subjects that were not used to generate the parcellation or train the classifier. The top two rows of Fig. 5 show the performance of the classifier in detecting each area (see the cortical areal classifier section in Methods). Importantly, the classifier aims to detect whole areas based on their multi-modal fingerprints, rather than detecting differences in areal features across paired areal boundaries as was done in the cross-validation analysis (Supplementary Fig. 6 in Supplementary Results and Discussion 1.2). The overall areal detection rate was 98.0% of all areas across all subjects for the 210P parcellation and training dataset (row 1) and 96.6% for the independent 210V validation dataset (row 2), indicating excellent overall performance of the areal classifier.

The areal classifier was used to generate probabilistic maps of each cortical area (illustrating residual variability in spatial location after MSMAll areal feature-based registration), and to assess the reproducibility of the parcellation in the independent 210V dataset. Rows 3 and 4 of Fig. 5 show strikingly similar probabilistic maps of 8 non-overlapping areas with differing degrees of spatial variability (V1, 4, RSC, MT, LIPv, TE1a, 46, and 10r) from the 210P and 210V groups. All probability maps were combined to produce a group maximum probability map (MPM), where the area with the highest probability at each vertex was found. Row 5 shows the original semi-automated parcellation borders, and row 6 compares the group MPM maps from 210P (blue) and 210V (red), with purple representing overlapping vertices. The borders in Row 6 are almost entirely purple, indicating very high reproducibility of the group MPM maps ($r = 0.965$, Dice = 0.960, see the cortical areal classifier section of Methods). This reproducibility is similar to that of the original group average feature maps discussed above. The correlation of the original semi-automated parcellation (row 5) with the 210P group MPM (row 6) was $r = 0.913$, Dice = 0.902, indicating that the classifier made modest adjustments to better fit the data. We predict there will be very high reproducibility of the parcellation across the rest of the ~1,100 subject HCP dataset. Example individual subject parcellations and their reproducibility based on repeated scan sessions are shown in Supplementary Fig. 11 of Supplementary Results and Discussion 1.4. The individual parcellations are reasonably reproducible (median $r = 0.77$, Dice = 0.72) but, unsurprisingly, not as reproducible as the group parcellations, which benefit from averaging across many subjects. Other analyses yield interesting information about the sizes of cortical areas in the group average and variability in areal size across individuals (Supplementary Results and Discussion 1.5).

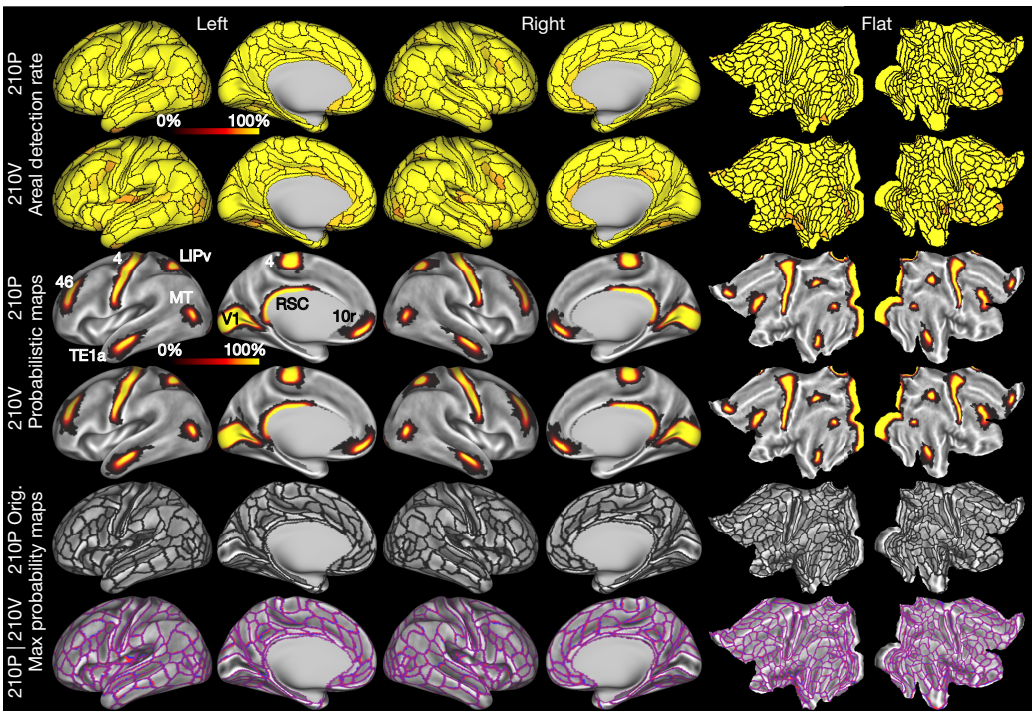


Figure 5 | Areal detection rates, probabilistic areas, and parcellation reproducibility. Rows 1 (210P) and 2 (210V) show the individual subject areal detection rates (see Methods section on cortical areal classifier) as parcellated maps. Most areas are yellow (100%), and the minimum detection rate across both rows was 73%. Rows 3 and 4 illustrate probabilistic maps of areas V1, 4, RSC, MT, LIPv, TE1a, 46, and 10r for the 210P (row 3) and 210V (row 4) groups. Row 5 shows the original parcellation derived from the semi-automated neuroanatomical approach. Row 6 shows the group MPM maps from 210P (blue), 210V (red), and their overlap (purple). Data at http://balsa.wustl.edu/WL8m.

Generalizing the classifier for future studies

In contrast to the semi-automated approach (above) where neuroanatomists chose the information to delineate and identify the 180 cortical areas in group-average data (see Supplementary Neuroanatomical Results), the areal classifier automatically determines (without human intervention) what information is most useful for delineating and identifying these cortical areas in individual subjects. As illustrated in Supplementary Figure 12 of Supplementary Results and Discussion 1.6, the areal classifier uses the task fMRI data least, perhaps because task fMRI feature maps are noisier in individual subjects than other feature maps and their information content is largely redundant with the resting state data[30]. This finding is important for generalizability of the areal classifier to other studies because replicating the customized, hour-long HCP task fMRI battery is unlikely to be feasible for most neuroimaging studies. Ideally, the areal classifier would be able to perform nearly as well relying only on architecture, connectivity, and topography. Accordingly, we trained the classifier again on the 210P dataset, but omitted the task fMRI-based feature maps. When trained this way, the classifier indeed performed nearly as well as when all features were used, detecting 97.6% of areas in 210P (versus 98.0% using all features) and 96.4% of areas in 210V (versus 96.6% using all features). Hence, we anticipate that the areal classifier will generalize to other studies that acquire the following core set of MRI images: high-resolution T1w and T2w; spin echo-based b0 field map; and extensive fMRI data acquired using 'multiband' pulse sequences to improve spatial and temporal resolution[7] (see Supplementary Results and Discussion 2.3 and 2.9). These are the same image acquisition requirements as the HCP's minimal preprocessing pipelines[5] and the MSMAll areal feature-based registration pipeline[14] (Supplementary Methods 2.4). Future studies adhering to these image acquisition guidelines will be able to use the unified framework of the HCP's analysis pipelines to automatically generate individualized parcellated analyses from unprocessed MRI images, a major advance over traditional neuroimaging methods that have often relied on comparisons with Brodmann's hand drawn parcellation published in 1909 (ref. 1).

Discussion

We have produced a population-based 180-area per hemisphere human cortical parcellation using exceptionally high quality multimodal data from hundreds of Human Connectome Project subjects aligned using an improved areal feature-based cross-subject alignment method (MSMAll). Inspired by an observer-independent post-mortem architectural parcellation approach[13], we developed a semi-automated neuroanatomical approach adapted to non-invasively acquired multi-modal MRI data. Although algorithms determined the final areal borders, the multi-modal data were carefully interpreted by neuroanatomists, the properties of each cortical area were documented, and each area was named in relation to the extant neuroanatomical literature (see Supplementary Neuroanatomical Results). A cross-validation showed that the areas forming the parcellation were robustly and statistically significantly different from their neighbours across multiple modalities. We identify this parcellation as HCP-MMP1.0 (Human Connectome Project Multi-Modal Parcellation version 1.0), making the version 1.0 designation because we anticipate future refinements as better data become available (see Supplementary Results and Discussion 2.1).

Unexpectedly, we discovered that despite improved intersubject alignment, some areas have atypical topological arrangements in some subjects, which we demonstrated for areas 55b, FEF, and PEF. We developed a fully automated method for parcellating individual subjects based on a machine learning classifier that can cope with this kind of individual variability. The areal classifier detected 96.6% of individual subject cortical areas in new subjects, including atypical areas, and replicated the group parcellation in an independent sample. Though we made extensive use of the HCP's specialized task fMRI battery when generating the parcellation, we showed that task fMRI data is not essential for future studies aiming to use the areal classifier to automatically define the cortical areas in their subjects. Instead, it suffices to acquire the same core set of MRI images needed for the rest of the HCP's software pipelines.

By generating a robust neuroanatomical map of human neocortical areas—a century-old aim of neuroscience—and providing methods for mapping these areas in any individual undergoing study with non-invasive neuroimaging, the present work represents a major advance relative to previous human cortical parcellations. The overall approach described here shows that we can produce sharp, reproducible brain images across multiple non-invasive neuroimaging modalities. We can generate a highly reproducible and generalizable cortical parcellation through state-of-the-art methods of data acquisition, preprocessing, and analysis designed to compensate for individual variability and thereby minimize blurring of images. These improvements, together with the new parcellation, make it desirable to use spatial localization methods that move beyond the traditional use of stereotaxic coordinates combined with Brodmann areal assignments to characterize centers of cortical activation in fMRI studies. From a neuroanatomical perspective, there has often been substantial uncertainty whether any two neuroimaging studies have found results in the same cortical areas or not. The situation is analogous to astronomy in which ground-based telescopes produced relatively blurry images of the sky before the advent of adaptive optics and space telescopes.

Many topics are discussed further in the Supplementary Results and Discussion 2.1–2.10 (for example, avenues for improving the parcellation and other issues left for future work, further discussion of the neuroscientific implications of our results, and additional datasets that could profitably be linked to our parcellation). As the topographic organization of higher cognitive areas becomes better understood, some parcels currently considered to be full areas may later be considered to be subareas of larger topographically organized cortical areas (analogous to somatotopic subregions of topographically organized sensory and motor areas illustrated in Supplementary Neuroanatomical Results 6). Though our use of multiple modalities probably mitigates this issue relative to traditional uni-modal parcellations, the extent to which the human multi-modal cortical parcellation may be revised along such lines remains a question for future work using the state of the art methods mentioned above (see Supplementary Results and Discussion 2.8).

The MSMAll registration and the areal classifier are or will soon be freely available on GitHub; the visualization tool Connectome Workbench is on http://humanconnectome.org; and the parcellation, data, and scenes for reproducing each of the figures are in the BALSA database[31]. These tools provide a neuroanatomical foundation, enabling the identification of cortical areas when reporting results or thinking about and discussing brain organization in relation to studies of human cognition, lifespan, and disease. Several additional interesting avenues of investigation are now open. The ability to discriminate individual differences in the location, size, and topology of cortical areas from differences in their activity or connectivity should facilitate the dissection of how each property is related to behaviour and genetic underpinnings, for example, in learning disabilities or those with distinctive cognitive traits. The ability to non-invasively and automatically delineate cortical areas in living subjects may have clinical implications, for example by providing neurosurgeons with detailed, individualized maps of the brains on which they operate. There are also important implications for our understanding of human cortical evolution. The dramatic expansion in neocortex along the human lineage occurred mainly in higher cognitive regions of lateral prefrontal, parietal, and temporal cortices[9,13,32,33]. Comparisons with nonhuman primates, including marmosets and macaques (both widely used in invasive studies), and great apes, may yield new insights regarding the emergence of new cortical areas and the divergences in areal functions, which collectively led to the cognitive capabilities that make us uniquely human as a species and as individuals.

Note added in proof: A related paper on the neuroimaging approach used by the Human Connectome Project may be found in ref. 34.

In addition, we note that FreeSurfer uses an algorithm to label gyri and sulci automatically in individual subjects based on manually generated training labels[35] that is similar in spirit to our areal classifier. Also, the FreeSurfer surface modelling noted in the Methods draws from methods summarized in ref. 36.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 12 November 2015; accepted 15 June 2016.
Published online 20 July 2016.

1. Brodmann, K. *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues* (J. A. Barth, 1909); *Brodmann's Localization in the Cerebral Cortex* (Smith Gordon, 1994) [transl. Garey, L.J.].
2. Felleman, D. J. & Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1,** 1–47 (1991).
3. Nieuwenhuys, R. The myeloarchitectonic studies on the human cerebral cortex of the Vogt–Vogt school, and their significance for the interpretation of functional neuroimaging data. *Brain Struct. Funct.* **218,** 303–352 (2013).
4. Van Essen, D. C., Glasser, M. F., Dierker, D. L., Harwell, J. & Coalson, T. Parcellations and hemispheric asymmetries of human cerebral cortex analyzed on surface-based atlases. *Cereb. Cortex* **22,** 2241–2262 (2012).
5. Glasser, M. F. *et al.* The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* **80,** 105–124 (2013).
6. Smith, S. M. *et al.* Resting-state fMRI in the Human Connectome Project. *Neuroimage* **80,** 144–168 (2013).
7. Uğurbil, K. *et al.* Pushing spatial and temporal resolution for functional and diffusion MRI in the Human Connectome Project. *Neuroimage* **80,** 80–104 (2013).
8. Van Essen, D. C. *et al.* The WU-Minn Human Connectome Project: an overview. *Neuroimage* **80,** 62–79 (2013).
9. Glasser, M. F., Goyal, M. S., Preuss, T. M., Raichle, M. E. & Van Essen, D. C. Trends and properties of human cerebral cortex: correlations with cortical myelin content. *Neuroimage* **93,** 165–175 (2014).
10. Glasser, M. F. & Van Essen, D. C. Mapping human cortical areas in vivo based on myelin content as revealed by T1- and T2-weighted MRI. *J. Neurosci.* **31,** 11597–11616 (2011).
11. Barch, D. M. *et al.* Function in the human connectome: task-fMRI and individual differences in behavior. *Neuroimage* **80,** 169–189 (2013).
12. Caspers, S., Eickhoff, S. B., Zilles, K. & Amunts, K. Microstructural grey matter parcellation and its relevance for connectome analyses. *Neuroimage* **80,** 18–26 (2013).
13. Schleicher, A., Amunts, K., Geyer, S., Morosan, P. & Zilles, K. Observer-independent method for microstructural parcellation of cerebral cortex: a quantitative approach to cytoarchitectonics. *Neuroimage* **9,** 165–177 (1999).
14. Robinson, E. C. *et al.* MSM: a new flexible framework for multimodal surface matching. *Neuroimage* **100,** 414–426 (2014).
15. Zilles, K. & Amunts, K. Centenary of Brodmann's map—conception and fate. *Nat. Rev. Neurosci.* **11,** 139–145 (2010).
16. Cohen, A. L. *et al.* Defining functional areas in individual human brains using resting functional connectivity MRI. *Neuroimage* **41,** 45–57 (2008).
17. Kolster, H., Peeters, R. & Orban, G. A. The retinotopic organization of the human middle temporal area MT/V5 and its cortical neighbors. *J. Neurosci.* **30,** 9801–9820 (2010).
18. Wang, L., Mruczek, R. E., Arcaro, M. J. & Kastner, S. Probabilistic maps of visual topography in human cortex. *Cereb. Cortex* **25,** 3911–3931 (2015).
19. Gordon, E. M. *et al.* Generation and evaluation of a cortical area parcellation from resting-state correlations. *Cereb. Cortex* **26,** 288–303 (2016).
20. Shen, X., Tokoglu, F., Papademetris, X. & Constable, R. T. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. *Neuroimage* **82,** 403–415 (2013).
21. Yeo, B. T. *et al.* The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106,** 1125–1165 (2011).
22. Hopf, A. Uber die Verteilung myeloarchitektonischer Merkmale in der Stirnhirnrinde beim Menschen. *J. Hirnforsch.* **2,** 311–333 (1956).
23. Van Essen, D. C. & Glasser, M. F. *In vivo* architectonics: a cortico-centric perspective. *Neuroimage* **93,** 157–164 (2014).
24. Olman, C. A. *et al.* Layer-specific fMRI reflects different neuronal computations at different depths in human V1. *PLoS One* **7,** e32536 (2012).
25. Polimeni, J. R., Fischl, B., Greve, D. N. & Wald, L. L. Laminar analysis of 7T BOLD using an imposed spatial activation pattern in human V1. *Neuroimage* **52,** 1334–1346 (2010).
26. Yacoub, E., Harel, N. & Ugurbil, K. High-field fMRI unveils orientation columns in humans. *Proc. Natl Acad. Sci. USA* **105,** 10607–10612 (2008).
27. Zimmermann, J. *et al.* Mapping the organization of axis of motion selective features in human area MT using high-field fMRI. *PLoS One* **6,** e28716 (2011).
28. Smith, S. M. *et al.* Functional connectomics from resting-state fMRI. *Trends Cogn. Sci.* **17,** 666–682 (2013
29. Hacker, C. D. *et al.* Resting state network estimation in individual subjects. *Neuroimage* **82,** 616–633 (2013).
30. Tavor, I. *et al.* Task-free MRI predicts individual differences in brain activity during task performance. *Science* **352,** 216–220 (2016).
31. Van Essen, D. C. *et al.* The brain analysis library of spatial maps and atlases (BALSA) database. *Neuroimage* http://dx.doi.org/10.1016/j.neuroimage.2016.04.002 (2016).
32. Hill, J. *et al.* A surface-based analysis of hemispheric asymmetries and folding of cerebral cortex in term-born human infants. *J. Neurosci.* **30,** 2268–2276 (2010).
33. Van Essen, D. C. & Dierker, D. L. Surface-based and probabilistic atlases of primate cerebral cortex. *Neuron* **56,** 209–225 (2007).
34. Glasser, M. F. *et al.* The Human Connectome Project's neuroimaging approach. *Nat. Neuroscience* (in press).
35. Fischl, B. *et al.* Automatically parcellating the human cerebral cortex. *Cereb. Cortex* **14,** 11–22 (2004).
36. Fischl, B. FreeSurfer. *NeuroImage* **62,** 774–781 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the members of the WU-Minn-Ox HCP Consortium for invaluable contributions to data acquisition, analysis, and sharing and E. Reid and S. Danker for assistance with preparing the manuscript. Supported by NIH F30 MH097312 (M.F.G.), ROIMH-60974 (D.C.V.E.), NIH F30 MH099877 (C.D.H.), the Human Connectome Project grant (1U54MH091657) from the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research, and the Wellcome Trust Strategic Award 098369/Z/12/Z (S.M.S., J.A., C.F.B., M.J.).

Author Contributions M.F.G. and D.C.V.E. designed the study and carried out the analyses. M.F.G., T.S.C., E.C.R., C.D.H., J.H., E.Y., K.U., J.A., C.F.B., M.J., and S.M.S. contributed novel methods. M.F.G., T.S.C., E.C.R., C.D.H., E.Y., J.A., C.F.B., M.J., S.M.S., and D.C.V.E. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.F.G. (glasserm@wustl.edu) or D.C.V.E. (vanessen@wustl.edu).

Reviewer Information *Nature* thanks R. Poldrack, F. Tong and T. Yeo for their contribution to the peer review of this work.

METHODS

Subjects and acquisition. A total of 449 young adult twins and non-twin siblings (ages 22–35) from the Human Connectome Project (HCP) were scanned according to the HCP's acquisition protocol[5–7]. The MRI acquisition included collecting T1w and T2w structural images, task-based and resting state-based fMRI images, diffusion-weighted images, and b0 field maps. Images were acquired at high spatial and temporal resolution on a customized Siemens 3 tesla (3T) scanner and with customized slice accelerated sequences for fMRI (see Supplementary Methods 1.1–1.2). All subjects from the HCP 500-subject data release (July, 2014) having complete fMRI sessions were included. They were divided into two independent groups of 210 subjects that shared no family members between them, together with a remaining group of 29 test (29T) subjects that shared family members with 210P but not 210V. The first group of subjects (210P, 130 females, 80 males) was used for creating the parcellation and training the areal classifier, which also made use of the 29T group to avoid overfitting. The second group of subjects (210V, 116 females, 94 males) was used only for statistical cross-validation of the parcellation, areal classifier detection rates in independent subjects, and group parcellation reproducibility measures. A test–retest group of 27 subjects scanned twice through the entire MRI protocol and independently processed through the HCP pipelines was used for individual subject reproducibility measures. Subject recruitment procedures and informed consent forms, including consent to share de-identified data, were approved by the Washington University institutional review board. Datasets were de-identified and are publicly shared on the ConnectomeDB database (https://db.humanconnectome.org).

Image preprocessing. Spatial image preprocessing (distortion correction and image alignment) was carried out using the HCP's spatial minimal preprocessing pipelines[5]. These pipelines maximize alignment across image modalities, minimize distortions relative to the subject's anatomical space, and minimize spatial smoothing (blurring) of the data. The data were projected into the 2mm standard CIFTI grayordinates space, which includes cortical grey matter surface vertices and subcortical grey matter voxels[5]. This offers substantial improvements in spatial localization over traditional volume-based analyses, enabling more accurate cross-subject and cross-study registrations and avoiding smoothing that mixes signals across differing tissue types or between nearby cortical folds. Additionally, we did minimal smoothing within the CIFTI grayordinates space to avoid mixing across areal borders prior to parcellation.

For cross-subject registration of the cerebral cortex, we used a two-stage process based on the multimodal surface matching (MSM) algorithm[14] (see Supplementary Methods 2.1–2.5). An initial 'gentle' stage, constrained only by cortical folding patterns (FreeSurfer's 'sulc' measure), was used to obtain approximate geographic alignment without overfitting the registration to folding patterns, which are not strongly correlated with cortical areas in many regions. Previously, we found that more aggressive folding-based registration (either MSM-based or FreeSurfer-based) slightly decreased cross-subject task-fMRI statistics, suggesting that aligning cortical folds too tightly actually reduces alignment of cortical areas[14]. A second, more aggressive stage used cortical areal features to bring areas into better alignment across subjects while avoiding neurobiologically implausible distortions or overfitting to noise in the data. The areal features used were myelin maps, resting state network maps computed with weighted regression (an improvement over dual regression[37] described in the Supplementary Methods 2.3) and resting state visuotopic maps (see Supplementary Methods 4.4). Areal distortion was measured by taking the log base-2 of the ratio of the registered spherical surface tile areas to the original spherical surface tile areas. The mean (across space) of the absolute value of the areal distortion averaged across subjects from both registration stages was 30% less than the standard FreeSurfer folding-based registration and the maximum (across space) of this measure was 54% less. Despite less overall distortion, the areal-feature-based registration delivers substantially more accurate registration of cortical areas than does FreeSurfer folding-based registration as judged by cross-subject task fMRI statistics, an areal feature that was not used to drive the registration[14]. Because MSM registration preserves topology and is relatively gentle (it does not tear or distort the cortical surface in neurobiologically implausible ways), it is unable to align some cortical areas in some subjects where the areal arrangement differs from the group average (see Supplementary Results and Discussion 1.3–1.4 for more details on atypical areas). Group average registration drift away from the gentle folding-based geographic alignment was removed from the surface registration[38] (see Supplementary Methods 2.5) to enable comparisons of this dataset with datasets registered using different areal features (for example, retinotopically defined areas). Group average registration drift is any consistent effect of the registration during template generation on the mean size, shape, or position of areas on the sphere (as opposed to the desired reductions in cross-subject variation). An obvious example is the 37% increase in average brain volume produced by registration to MNI space[4]. Uncorrected drifts during surface template generation can cause apparent changes in cortical areal size, shape, and position when comparing across studies.

Resting state fMRI data were denoised for spatially specific temporal artefacts (for example, subject movement, cardiac pulsation, and scanner artefacts) using the ICA+FIX approach, which includes detrending the data and aggressively regressing out 24 movement parameters[39,40]. We avoided regressing out the 'global signal' (mean grey-matter time course) from our data because preliminary analyses showed that this step shifted putative connectivity-based areal boundaries so that they lined up less well with other modalities, likely because of the strong areal specificity of the residual global signal after ICA+FIX clean up. Task fMRI data were temporally filtered using a high pass filter. More details on resting state and task fMRI temporal preprocessing are described in the Supplementary Methods 1.6–1.8. Substantial spatial smoothing was avoided for both datasets, and all images were intensity normalized to account for the receive coil sensitivity field. Artefact maps of large vein effects, fMRI gradient echo signal loss, and surface curvature were computed as described in Supplementary Methods 1.9.

Modalities for parcellation. The multi-modal cortical parcellation used information related to the four areal properties of architecture, function, connectivity, and topography[2]. Architecture was measured using T1w/T2w myelin content maps plus cortical thickness maps with surface curvature regressed out[5,9,10] (Supplementary Methods 1.5). Function was measured using task-fMRI responses to 7 tasks in 86 task contrasts (47 unique; 39 were sign-reversed contrasts). Effect size maps (beta maps) after correction for the receive field were used instead of *Z* statistic maps because we were interested in regional differences in the magnitude of the BOLD (blood oxygen level dependent) signal change induced by the tasks, rather than differences in the significance of the BOLD signal change. Functional connectivity was measured using pairwise Pearson correlation of the denoised resting state time series of each pair of grayordinates. Topographic organization was explored using resting state time series in visual cortex, with spatial regressors representing polar angle and eccentricity patterns in area V1 combined with a modified 'dual-regression-like' approach that weights each surface vertex according to the cortical surface area that it represents (see Supplementary Methods 4.4). The semi-automated multi-modal parcellation was generated using group average data for all of these modalities from the 210P group of subjects (see Supplementary Methods 3.1–3.3 for details on how the group averages were created for each modality). The reproducibility of these group average maps was assessed by correlating the spatial maps for the 210P and 210V groups (see Supplementary Results and Discussion 1.1).

The gradient-based parcellation approach. Classically, cortical areas have been defined based on sharp changes in one or more of the areal properties of architecture, function, connectivity, and topography. Traditionally, this relied heavily on visual inspection, until more objective and quantitative approaches became available[12,13]. One highly successful approach to post-mortem architectural parcellation involves computing a dissimilarity metric, (the Mahalanobis distance) between neighbouring feature profiles generated from segmented histological images and testing for statistically significant and large spikes in dissimilarity that indicate putative areal boundaries. For *in vivo* data, a similarly powerful approach involves taking the first derivative (the spatial gradient) of a measure of interest along cortical surface and using the gradient magnitude to objectively identify locations where the measure is changing rapidly. One can then draw putative areal boundaries along the resulting gradient ridges[10,16,19]. Here we combined elements of both approaches in a multi-modal context to generate semi-automatically drawn areal borders that were then evaluated statistically. Gradients were computed for architectural, functional, connectivity, and topographic modalities (see Supplementary Methods 4.1–4.4).

To incorporate expert knowledge and priors from the neuroanatomical literature into the parcellation process, the neuroanatomists (authors M.F.G and D.C.V.E.) evaluated the multi-modal neuroimaging data and its gradients to define initial areal borders based on the following criteria. (1) Presence of a co-localized gradient ridge in at least two independent modalities was taken as strong evidence of an areal border, and the vast majority of areal borders satisfied this criterion. (2) Presence of corresponding gradients in the left and right hemispheres provided further evidence for a genuine areal border. For the vast majority of borders, the same modalities yielded robust gradients in both hemispheres. We did not find strong evidence for an area present in one hemisphere that was absent in the other (though a few areas show hemispheric asymmetries in their functional 'signature' and/or in their spatial relationships with neighbouring areas). (3) We ignored gradients clearly attributable to imaging artefacts (see Supplementary Neuroanatomical Results for details). (4) Cortex on opposite sides of the border needed to differ robustly and significantly in the areal features used to delineate the border. (5) Confidence was increased if prior literature described a corresponding areal border. (6) Early runs of a supervised machine learning algorithm (see the

cortical areal classifier section of the Methods below) needed to be able to learn to distinguish each cortical area from its neighbours in a large majority of individual subjects based on individual subject multi-modal features (the early runs were only done using 210P and 29T, keeping 210V independent for later analyses). After the neuroanatomists delineated the initial areal borders and chose the important areal features that defined them, an automated algorithm then optimized the border placement so that it followed the most probable path based on the chosen areal features (see Supplementary Methods 5.1–5.3). The Supplementary Neuroanatomical Results documents the information that was used to distinguish each of the 180 areas from its neighbours.

The neuroanatomists named areas based on previous parcellations whenever a reasonable match to the literature could be made. In some cases, areal identification was based on the similarity of the area's properties relative to previously reported areas (for example, area 4, primary motor cortex, is known to be heavily myelinated and thick; area V2 has a mirror-image visuotopic map relative to neighbouring area V1). In most cases, however, the information used to describe previous cortical areas (for example, cytoarchitecture) was not available in the HCP data, and areal identification mainly reflected spatial correspondences relative to cortical folding patterns (if reliable for that region of cortex) or spatial relationships between neighbouring cortical areas. The strongest evidence for areal identification came from studies that provided surface-based probabilistic or maximum probability maps, ideally also registered using areal features and dedrifting of templates[38]. In these cases, we directly compared these data with our data and show the degree of overlap in the Supplementary Neuroanatomical Results. When such data were unavailable, we used published information to the degree feasible (see Supplementary Methods 5.3 for limitations of non-surface-based/not publicly available data) to make areal identifications or to describe new areas that had not previously been identified. The information used to name each cortical area is described in the Supplementary Neuroanatomical Results.

Statistical cross validation of the multi-modal parcellation. Once the parcellation has been created, parcellated representations of data from each modality can be generated using either the group parcellation or the individual subject parcellations. For the statistical cross-validation, we created parcellated myelin, cortical thickness, task fMRI, and resting state functional connectivity datasets using the semi-automated multimodal group parcellation (see Supplementary Methods 7.1). For myelin and cortical thickness, we simply averaged the values of the dense individual subject maps within each area. For task fMRI, we averaged the time series within each area prior to computing task statistics (to benefit from the SNR improvements of parcellation demonstrated in Fig. 4e). For the same reason, we averaged resting state time series within each parcel prior to computing functional connectivity to form a parcellated functional connectome.

For each pair of areas that shared a border in the parcellation, we computed a paired samples two-tailed t-test across subjects on these parcellated data for each feature (ignoring tests that involved the diagonal in the resting state parcellated functional connectome). We thresholded these tests at the Bonferroni-corrected significance level of $P < 9 \times 10^{-8}$ (number of area pairs across both hemispheres (1,050) × number of features (266) × number of tails (2) × 0.05) and an effect size threshold of Cohen's $d > 1$. We grouped the features into 4 independent categories (cortical thickness, myelin, task fMRI, and resting state fMRI) to determine for each area pair whether it showed robust and statistically significant differences across multiple modalities. For more details, see Supplementary Methods 7.2.

The cortical areal classifier. We used a supervised machine learning classifier to automatically delineate and identify each cortical area from its neighbours across a large majority of individual subjects based on multi-modal information. Besides validating the robustness of the parcellation, this provides useful information about each individual subject's parcellation, along with an approach to generalizing the parcellation to other datasets. To automatically parcellate individual subjects, we adapted the multi-layer perceptron used by ref. 29 to delineate and identify seven resting state networks more accurately than simpler linear methods including dual regression. We used the multi-layer perceptron to classify all 180 areas in our parcellation using multi-modal feature maps and relied on two neuroanatomically sensible assumptions to simplify the problem. (1) After areal feature-based registration (MSMAll), we assumed that each cortical area was approximately in the same general location across subjects (for example, we don't expect to find V1 outside the occipital lobe). This also means that we consider widely separated regions having similar multi-modal areal fingerprints to be distinct cortical areas even if they have similar architecture, coactivation in functional tasks, and belong to the same resting state network. These assumptions allowed us to reduce the overall classification problem to a set of 180 classification problems per hemisphere, each involving discrimination of one area from the areas around it. (2) Also, instead of classifying each area from all of its neighbours specifically (one class for the area plus one class for each neighbouring area), we set up the problem as a binary classification (the most robust kind of classification problem), classifying each area from all of the surrounding cortex as a single alternate class. This surrounding cortex represents a 'searchlight' for the area, and this searchlight was the group parcel location plus a 30 mm radius surrounding the group parcel in all directions across the surface (meaning that for a 10 mm circular area, the searchlight would be a circle of 70 mm in diameter, still a quite large region of cortex). The 30 mm radius (geodesic distance computed on the group average mid-thickness surface corrected for vertex area loss due to averaging) was chosen because it easily encompassed the individual variation in area 55b in the 210P group (55b approaches a worst case because it is a relatively small and highly variable cortical area). The training labels were the group area from the semi-automated parcellation (class 1), and the remaining cortex in the searchlight (class 2).

The features used by the classifier covered the same set of modalities used for the original parcellation, including architectural measures of myelin and cortical thickness with curvature regressed out; task fMRI maps (redundant information was reduced and SNR increased with a $d = 20$ ICA-decomposition run on the task contrast beta maps, see Supplementary Methods 6.4); the 77 surface-related resting state fMRI network maps computed on individual subjects using weighted regression from an overall $d = 137$ group ICA; five visuotopic topographic maps transformed into a format interpretable by the classifier; and maps of artefacts that the classifier used to interpret differences in areal features due to artefactual effects (see Supplementary Methods 6.3–6.5 for further description of each modality's classifier features). These 112 multi-modal feature maps were generated for each vertex in each of the 449 subjects and the 27 repeated subjects, with each hemisphere processed separately. Other than the 30 mm radius searchlight region of interest (ROI), the classifier has no spatial concept of where the area should be (it operates independently on each vertex and only knows what the area's fingerprint looks like in the feature space). Consequently, special consideration was given to the spatial visuotopic patterns, which were transformed into maps whose values reflected the alternating mirror symmetric organization of visual areas (that is, maps whose values reflect the orientation of the visuotopic gradient vector relative to the vector that points 'geodesically' towards V1, see Supplementary Methods 6.5).

The classifier analyses were conducted using a standard machine learning train/test/validation approach. The classifier was trained using the 210P subjects and tested against overfitting using the extra 29T subjects. The 210V subjects were used as the validation sample, and thus were not involved in the classifier training, testing, or the parcellation itself, and also shared no family relationships with the 210P or the 29T groups. A short initial run of the classifier was used to identify features that the classifier was particularly sensitive to for each area (see below and Supplementary Methods 6.6). These features were compared in each individual subject with the group average pattern to exclude subjects that were potentially misaligned with the typical subject in this region (and hence for which the group defined training labels were likely inaccurate). This area-specific set of subjects in the 210P and 29T groups were excluded from the final classifier training of each area. The classifier's output (ranging from 0 to 1) represents the likelihood that a given vertex in a subject is part of the area being classified or part of the surrounding cortex of the searchlight. Once the classifier training weights have been generated, it is possible to classify any subject who has the 112 multi-modal maps computed, including those whose areas are misaligned with the group (see Supplementary Results and Discussion 1.4).

The trained classifier was applied to the 449 subjects and 27 repeat subjects to generate individual subject likelihood maps for each of the 180 areas in each hemisphere. These probability maps were combined by finding the largest probability for each vertex and then regularized within local neighbourhoods (see Supplementary Methods 6.7) to make an individual subject 'winner-take-all' parcellation. An area was considered to have been detected in a subject for the purposes of the areal detection measures (the overall classifier areal detection rate and the maps of areal detection rate for each area) if its size was between 1/3 and 3 times the size of the original population-based parcel (a pragmatic threshold chosen prior to performing the analysis that tolerates modestly greater neuroanatomical variability across subjects than the empirical range reported in cytoarchitectonic studies[41,42]). Probabilistic maps of each area were then created by separately averaging the individual subject winner-take-all parcellation areas for the 210P and 210V subject groups. A group maximum probability map (MPM) parcellation was then created by assigning the identity of the maximum areal probability to each vertex. The reproducibility of the parcellation was assessed by correlating these two MPM maps and by computing a Dice coefficient. In both cases the parcellation was first turned into 180 concatenated binary ROIs per hemisphere (each area was represented by a separate map, ~30,000 vertices per hemisphere, with ones for all vertices inside the area and zeros

for all vertices outside). The reproducibility of the individual subject hard parcellation maps was assessed similarly. For more details, see Supplementary Methods 7.2.

Multi-modal areal fingerprints learned by the classifier were visualized using a classifier sensitivity metric. This metric was the partial derivative with respect to each feature of each area multiplied by the gradient magnitude of the feature (see Supplementary Methods 6.8). The measure indicates which areal features the classifier finds most informative when classifying a given area and whether increases or decreases in the value of the feature make the area more likely to be present. The sensitivity metric can be visualized both at the dense (vertex-wise) level for each feature and each area, or summarized at a parcel level. For each feature, the sensitivity metric was summarized at the parcel level by taking the maximum absolute value of the metric (finding the border where the feature was most influential) and using this maximum to represent the area in a parcellated or a matrix view, as shown in Supplementary Fig. 12 of Supplementary Results and Discussion 1.6.

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

37. Filippini, N. *et al.* Distinct patterns of brain activity in young carriers of the *APOE-ε4* allele. *Proc. Natl Acad. Sci. USA* **106,** 7209–7214 (2009).
38. Abdollahi, R. O. *et al.* Correspondences between retinotopic areas and myelin maps in human visual cortex. *Neuroimage* **99,** 509–524 (2014).
39. Griffanti, L. *et al.* ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. *Neuroimage* **95,** 232–247 (2014).
40. Salimi-Khorshidi, G. *et al.* Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage* **90,** 449–468 (2014).
41. Caspers, S. *et al.* The human inferior parietal lobule in stereotaxic space. *Brain Struct. Funct.* **212,** 481–495 (2008).
42. Malikovic, A. *et al.* Cytoarchitectonic analysis of the human extrastriate cortex in the region of V5/MT+: a probabilistic, stereotaxic map of area hOc5. *Cereb. Cortex* **17,** 562–574 (2007).

ARTICLE

doi:10.1038/nature19068

SAR11 bacteria linked to ocean anoxia and nitrogen loss

Despina Tsementzi[1], Jieying Wu[2], Samuel Deutsch[3], Sangeeta Nath[3], Luis M. Rodriguez-R[2], Andrew S. Burns[2], Piyush Ranjan[2], Neha Sarode[2], Rex R. Malmstrom[3], Cory C. Padilla[2], Benjamin K. Stone[4], Laura A. Bristow[5], Morten Larsen[6], Jennifer B. Glass[7], Bo Thamdrup[6], Tanja Woyke[3], Konstantinos T. Konstantinidis[1,2] & Frank J. Stewart[2]

Bacteria of the SAR11 clade constitute up to one half of all microbial cells in the oxygen-rich surface ocean. SAR11 bacteria are also abundant in oxygen minimum zones (OMZs), where oxygen falls below detection and anaerobic microbes have vital roles in converting bioavailable nitrogen to N_2 gas. Anaerobic metabolism has not yet been observed in SAR11, and it remains unknown how these bacteria contribute to OMZ biogeochemical cycling. Here, genomic analysis of single cells from the world's largest OMZ revealed previously uncharacterized SAR11 lineages with adaptations for life without oxygen, including genes for respiratory nitrate reductases (Nar). SAR11 *nar* genes were experimentally verified to encode proteins catalysing the nitrite-producing first step of denitrification and constituted ~40% of OMZ *nar* transcripts, with transcription peaking in the anoxic zone of maximum nitrate reduction activity. These results link SAR11 to pathways of ocean nitrogen loss, redefining the ecological niche of Earth's most abundant organismal group.

Alphaproteobacteria of the SAR11 clade form one of the most ecologically dominant organism groups on the planet, representing up to half of the total microbial community in the oxygen-rich surface ocean[1–5]. All characterized SAR11 isolates, including the globally ubiquitous *Candidatus* Pelagibacter genus, are aerobic heterotrophs adapted for scavenging dissolved organic carbon and nutrients under the oligotrophic conditions of the open ocean[6–9]. Gene-based surveys have also revealed diverse SAR11 lineages at high abundance in the deep waters of the meso- and bathypelagic realms[10–13]. However, the functional properties that distinguish SAR11 bacteria living in distinct ocean regions remain unclear. All known SAR11 genomes are small (typically less than 1.5 megabase pairs (Mb)), with genomic streamlining as a potential adaptation to the nutrient-limiting conditions of the open ocean[11]. It has been hypothesized that adaptations in SAR11 do not involve large variations in gene content[6,8], suggesting that the contribution of SAR11 to ocean biogeochemistry is primarily through its role in aerobic oxidation of organic carbon.

Although genetic or biochemical evidence of anaerobic metabolism has not been reported for SAR11, high abundances of SAR11-related genes have been detected under anoxic conditions in marine OMZs. Permanent OMZs extend over ~8% of the oceanic surface area (oxygen (O_2) < 20 μM)[14], with the largest and most intense OMZs in upwelling regions of the Eastern Pacific. In the cores of these regions, microbial respiration of high surface primary production combines with low ventilation to deplete O_2 from mid-water depths, resulting in O_2 concentrations below detection (~10 nM) over a major portion (~100–700 m) of the water column[15]. In the absence of O_2, respiratory nitrate (NO_3^-) reduction to nitrite (NO_2^-) becomes the dominant process for organic matter oxidation[16], with respiratory Nar proteins being among the most abundant and highly expressed enzymes in OMZs[17–19]. NO_3^- respiration results in a substantial accumulation of NO_2^- in OMZs, often to micromolar concentrations[20]. This NO_2^- pool is actively cycled through NO_2^--consuming microbial metabolisms, notably the anaerobic processes of denitrification and anaerobic ammonium oxidation (anammox)[21,22], which together in OMZs account for 30–50% of the loss of bioavailable nitrogen from the ocean as either gaseous dinitrogen (N_2) or nitrous oxide (N_2O)[21,22]. Surprisingly, SAR11 bacteria are often the most abundant organisms in the NO_2^--enriched N-loss zone of OMZs, where O_2 is undetectable, representing ~20% (range: 10–40%) of all 16S ribosomal RNA genes and protein-coding metagenome sequences in the 0.2–1.6 μm biomass fraction[18,19,23,24]. Such high abundances imply that SAR11 make up a substantial fraction of the OMZ community and raise the question of the role of SAR11 in OMZ biogeochemistry.

We analysed single amplified genomes (SAGs) to identify the metabolic basis for the dominance of SAR11 in anoxic OMZs. We focused on SAR11 SAGs obtained from the Eastern Tropical North Pacific (ETNP) OMZ off Mexico, the world's largest OMZ, accounting for 41% of global OMZ surface area[14] (Fig. 1a). O_2 concentration at this site declined from ~200 μM at the surface to ~400 nM at the bottom of the oxycline (30–85 m) and was typically at or below the detection limit (~10 nM) from ~90 m to 700 m. At the time of sample collection, NO_3^- reduction rates increased with depth into the OMZ, peaking at ~9.5 nM N d^{-1} at 300 m (ref. 19), paralleling an increase in the abundance of sequences encoding Nar-type NO_3^- reductases in coupled metagenomes and metatranscriptomes (Fig. 1c). In contrast, aerobic NO_2^- oxidation peaked at 100 m (260 nM N d^{-1}), where trace O_2 was available and NO_2^- was abundant, before declining 20-fold with depth into the OMZ (Fig. 1c). However, NO_2^- oxidation rates are probably overestimated due to slight O_2 contamination in incubations[20]. These data highlight a transition to anoxia within the ETNP OMZ[15,19], with *in situ* O_2 concentration at least an order of magnitude lower than the inhibitory threshold for NO_3^- reduction, denitrification and anammox[25,26], consistent with micromolar accumulations of NO_2^- from NO_3^- reduction in this zone.

[1]School of Civil and Environmental Engineering, Georgia Institute of Technology, Ford Environmental Science & Technology Building, 311 Ferst Drive, Atlanta, Georgia 30332, USA. [2]School of Biological Sciences, Georgia Institute of Technology, Ford Environmental Sciences & Technology Building, 311 Ferst Drive, Atlanta, Georgia 30332, USA. [3]Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598, USA. [4]Department of Biology, Bowdoin College, 255 Maine St, Brunswick, Maine 04011, USA. [5]Biochemistry Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany. [6]Department of Biology and Nordic Center for Earth Evolution (NordCEE), University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark. [7]School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Ford Environmental Sciences & Technology Building, 311 Ferst Drive, Atlanta, Georgia 30332, USA.

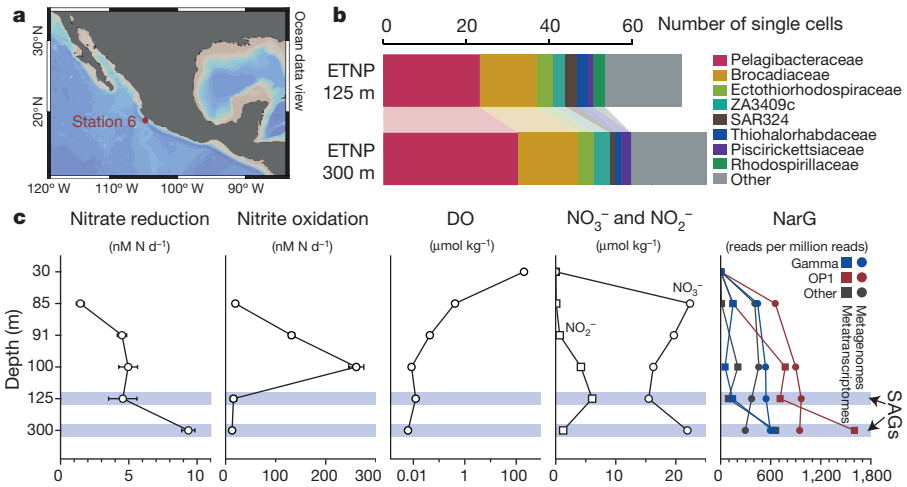


Figure 1 | Site description and phylogenetic affiliation of single cells. a, Location of station 6 (red) in the ETNP from which samples were obtained. Map was created with Ocean Data View (http://odv.awi.de). **b**, Taxonomic classification of sorted single cells, based on their 16S rRNA genes. **c**, Nitrate reduction and nitrite oxidation rates relative to dissolved O_2 (DO), nitrate and nitrite concentrations and *narG* read abundance in metagenomes and metatranscriptomes. Error bars represent standard error of the mean (s.e.m.) from triplicate measurements. Note that a $\log_{10}$ scale is used for the DO plot and that 0.01 μmol kg^{-1} represents the detection limit of the STOX sensor oxygen data presented here. DO at 300 m was below the detection limit.

Diverse SAR11 SAGs from anoxic waters

Samples for SAG analysis were obtained from two depths in the anoxic zone: at 125 m at the NO_2^- maximum (6 μM), and at 300 m in the core of the NO_3^- reduction zone. Single prokaryotic cells were isolated by fluorescence-activated cell sorting, subjected to genome amplification[27], and screened by 16S rRNA gene fragment (470 bp) polymerase chain reaction (PCR) and Sanger sequencing. From this screen, 23% and 32% of SAGs from 125 m and 300 m, respectively, were confidently assigned to the SAR11 family Pelagibacteraceae (Fig. 1b), thus confirming the substantial numerical abundance of SAR11 in the OMZ. From this SAR11 subset, 10 SAGs from 125 m and 12 SAGs from 300 m were randomly selected for shotgun sequencing (Illumina), along with 5 technical control SAR11 SAGs from the oxic surface waters of the Gulf of Mexico (GoM). After sequencing, quality filtering and assembly, a total of 19 SAGs were used for analysis: 15 OMZ SAGs (5 from 125 m, 10 from 300 m) and 4 GoM control SAGs (Supplementary Table 1). These genomes exhibited varying levels of completeness (~2–90%; average 30%) and no detectable contamination (Extended Data Fig. 1), as assessed by the presence of single-copy housekeeping genes[28,29], 16S rRNA gene identities, and the taxonomic assignment of SAG contigs (Supplementary Tables 1, 2 and Supplementary Discussion).

The identified SAGs represented a diverse and novel SAR11 community in the OMZ. Phylogenetic reconstructions based on either 16S rRNA genes or single-copy housekeeping proteins placed the 19 SAGs in 5 subclades of SAR11 (Fig. 2a). Average amino acid identity (AAI) comparisons among all available SAR11 genomes (Supplementary Table 3) further corroborated this classification, placing: (1) seven OMZ SAGs within the previously uncharacterized deep-branching monophyletic group of subclade IIa (hereafter designated subclade IIa.A), distinct (>5% 16S divergence) from SAG HIMB058 from the tropical North Pacific (hereafter designated subclade IIa.B); (2) three OMZ SAGs within the deep-branching subclade IIb; (3) two OMZ SAGs within subclade Ic, which includes recently described SAGs from the bathypelagic ocean[6]; (4) two OMZ and all four GoM surface SAGs within subclade Ib, which thus far lacks genome representatives; and (5) OMZ SAG A7 as most closely related to HIMB59, a member of the divergent SAR11 subclade V[8,30,31]. Note that the exact placement of subclade V in the SAR11 phylogeny is unstable depending on the marker gene and outgroup used[32,33]. The average estimated genome size of OMZ SAGs was 1.33 Mb (Supplementary Table 1), consistent with prior reports of genome streamlining in SAR11.

OMZ SAR11 abundance peaks under oxygen depletion

To estimate the *in situ* abundance and activity of OMZ SAR11, metagenome and metatranscriptome reads from OMZ sites and from diverse oxic ocean regions (Supplementary Table 4) were recruited to 39 available SAR11 genomes (Supplementary Table 1). Metagenomic read recruitment, performed essentially as described previously[34], showed that each OMZ SAR11 subclade represents a sequence-discrete (and hence tractable) population (Supplementary Discussion), but with each population encompassing substantial intra-population variation (~92–100% average nucleotide identity between members of the population versus <90% between populations), as well as gene content variability (Extended Data Fig. 2). We therefore estimated SAR11 abundance at the subclade level, based on the average coverage of 507 genes shared between genomes from all SAR11 subclades. On the basis of this analysis, SAR11 subclades Ic, IIa.A and IIb together comprised about 10–30% of the bacterial community in ETNP and ETSP metagenomes and metatranscriptomes from depths with undetectable O_2 (Fig. 2b, c), consistent with the high abundance of SAR11 in the pool of cells sorted for SAG analysis (Fig. 1b). Subclade IIa.A, composed exclusively of seven SAGs from this study, was particularly abundant, making up to 15% of the community in anoxic samples. All OMZ subclades were absent from or much less abundant (<5%) in metagenomes from oxic sites, including those from above the ETNP OMZ (Fig. 2b). Together, these results identify newly described SAR11 subclades whose distribution is linked to an oxygen-depleted niche.

Metabolic adaptations to low oxygen in SAR11 genomes

OMZ and GoM SAGs were then analysed for evidence of microaerobic or anaerobic metabolism. Surprisingly, in 8 of the 15 OMZ SAGs, belonging to SAR11 subclades Ic, IIa.A, IIb and V, protein family-based classification detected genes encoding the respiratory Nar of the DMSO reductase superfamily (Fig. 2a). Evidence of a complete canonical *nar* operon (*narGHJI*)—encoding the α subunit that catalyses NO_3^- reduction to NO_2^- (NarG), the iron–sulfur-containing β subunit (NarH) that transfers electrons to the molybdenum cofactor of NarG, the transmembrane cytochrome *b*-like γ subunit (NarI) involved in electron transfer from membrane quinols to NarH, and the NarJ chaperone involved in enzyme formation—was found within a single assembled contig in four SAGs (A6, E4, D9, A7), while partial *narG* and *narH* fragments were identified in another four SAGs (Extended Data Fig. 3). In all SAR11 SAGs containing *nar* on a contig, we identified other genes upstream or downstream on the same contig taxonomically assigned to SAR11 reference genomes (Supplementary Table 5 and Supplementary Discussion), further confirming the association of *nar* with SAR11. Genes encoding the NO_3^-/ NO_2^- transporter NarK and proteins for biosynthesis of the essential molybdenum cofactor (*moeA*, *mobA*) were also identified in eight and five of the SAGs, respectively (Supplementary Table 1). In only four of the fifteen OMZ SAGs were *nar* or cofactor synthesis genes not detected, presumably due to sequencing gaps (completeness of these SAGs: 4–20%; Supplementary Table 1). In contrast, these genes were not detected in any of the four control SAGs from the oxic GoM, despite high completeness of those

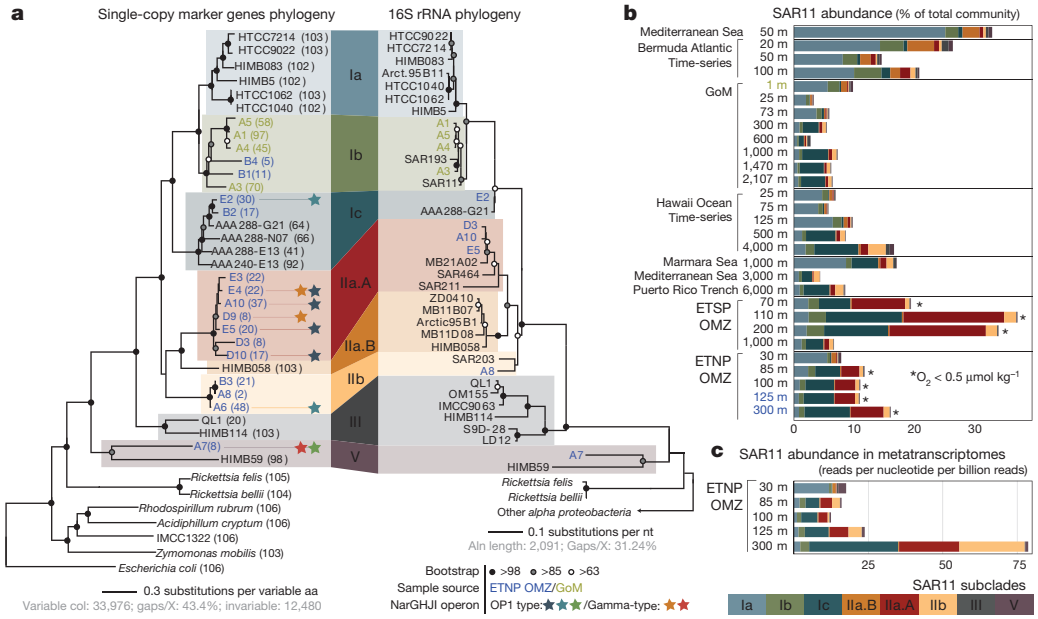


Figure 2 | Diversity, abundance and transcription of nitrate-reducing SAR11. a, Maximum likelihood phylogeny based on the concatenated alignment of single copy housekeeping (left) and 16S rRNA (right) genes in SAGs from this study, SAR11 and representative alphaproteobacterial genomes. Values in parentheses denote the number of housekeeping genes used per genome. For the 16S-based tree, only full-length sequences from the genomes in the left tree were included. Star symbols of the same colour represent closely related *narG* genes (>97% amino acid identity), encoding the catalytic subunit of the respiratory nitrate reductase of the DMSO family. aa, amino acid; nt, nucleotide. **b**, Abundance of SAR11 subclades (left) in selected oceanic metagenomes. Note that the major *nar*-encoding clade IIa.A peaks in abundance at oxygen-depleted OMZ depths. Data set descriptions are available in Supplementary Table 1. **c**, Normalized average coverage of SAR11 subclades in ETNP metatranscriptomes. Transcription by *nar*-encoding lineages increases from the base of the oxycline (85 m) to spike at the OMZ core (300 m), but is negligible in the overlying oxic zone (30 m).

genomes (average 61%). Genes encoding for downstream steps of denitrification or other dissimilatory anaerobic metabolisms were not found in any of the SAGs. However, in contrast to all previously analysed SAR11 genomes, three of the OMZ SAGs, all from subclade IIa.A, also contained genes encoding high-affinity O_2-using *bd*-type terminal oxidases (Supplementary Table 1). Compared with the *coxI*-type oxidases present in all known SAR11 genomes, including the OMZ SAGs analysed here, *bd*-type oxidases have a much higher affinity for O_2 (3–8 nM; Supplementary Discussion), suggesting a potential for microaerobic respiration by OMZ SAR11. These results provide the first indication of adaptation to low oxygen in SAR11 and the ability to respire NO_3^- to NO_2^- in the absence of oxygen, consistent with the distribution of these bacteria in the OMZ water column.

Multiple divergent Nar proteins in OMZ SAR11

Phylogenetic placement of all identified *narG* and *narH* genes and partial fragments revealed two divergent *nar* variants in OMZ SAGs (Fig. 3a and Extended Data Fig. 3): (1) an 'OP1 type' in which all four *nar* genes and an upstream cytochrome *c* protein were most similar (56–78% amino acid identity) to homologues from '*Candidatus* Acetothermus autotrophicum' (Supplementary Table 5), a putative anaerobic acetogen of the candidate bacterial phylum OP1 (ref. 35); and (2) a 'Gamma-type' variant most similar (51–78% identity) to Nar from a denitrifying Gammaproteobacteria endosymbiont (*Ca.* Vesicomyosocius okutanii strain HA)[36]. At least two of the OMZ SAR11 SAGs from subclade IIa.A, as well as SAG A7 from subclade V, encoded both OP1- and Gamma-type *nar* variants, suggesting that divergent *nar* copies (~42% amino acid identity) co-occur in the same genome (Supplementary Discussion). Multiple *nar* operons per genome have been reported for diverse bacteria and are hypothesized to be related to adaptation to different oxygen conditions, with one variant constitutively expressed at low baseline levels[37–39]. For both OP1- and Gamma-type variants, the sequence divergence among recovered sequences was consistent with the phylogenetic placement of the SAGs. For example, OP1-type *narG* fragments represented three distinct 97% amino acid identity clusters (Fig. 2a). Sequences from clade IIa.A SAGs fell within the same cluster, sharing ~96.5% identity with sequences of the closely related Ic and IIb subclades, and ~90% with sequences from the more distant A7 SAG (Extended Data Fig. 3). This pattern suggests diversification of *nar* operons in parallel with its genomic background, and also confirms that these sequences are not a systemic contaminant (Supplementary Discussion).

Biochemical characterization of SAR11 Nar

We sought to characterize further the biochemical function of SAR11 *nar* genes. Phylogenetic reconstruction based on 392 proteins of the diverse DMSO superfamily revealed that both OP1- and Gamma-type NarG fall within the clade of membrane-bound cytoplasm-oriented Nar and NO_2^- oxidoreductases (Nxr), and were most closely related to Nar from known NO_3^--reducing bacteria (Fig. 3a)[40]. The lack of a TAT peptide motif at the N terminus corroborated the probable cytoplasmic orientation of the NarG active site[41], similar to experimentally verified Nar in *Escherichia coli*[42]. Additionally, the identified NarG sequences contain diagnostic functional domains found in NarG but not in other oxidoreductases of the DMSO reductase superfamily (Extended Data Fig. 4)[40].

To verify NO_3^- reduction potential in SAR11, we introduced full-length SAR11 *nar* operons into a NO_3^- reductase-deficient *E. coli* mutant and tested for enzyme activity. The Gamma-type *nar* operon was successfully expressed in *E. coli*, yielding Nar proteins of the predicted size range and enabling growth of the mutant under anoxic conditions in the presence of NO_3^-, coupled with simultaneous NO_3^- reduction to NO_2^- (Extended Data Fig. 5), thereby providing direct evidence for the function of this enzyme *in vivo*. The OP1-type operon did not reverse the *E. coli* mutant phenotype, presumably due to the much greater divergence of this variant from the *E. coli nar* operon. Given the high similarity of Nar and Nxr protein sequences[43–45], and the reversibility of the NO_3^- reduction reaction, it is possible that either or both OP1- and Gamma-type proteins could also function *in situ* to oxidize NO_2^- aerobically. Although it is enticing, this possibility is

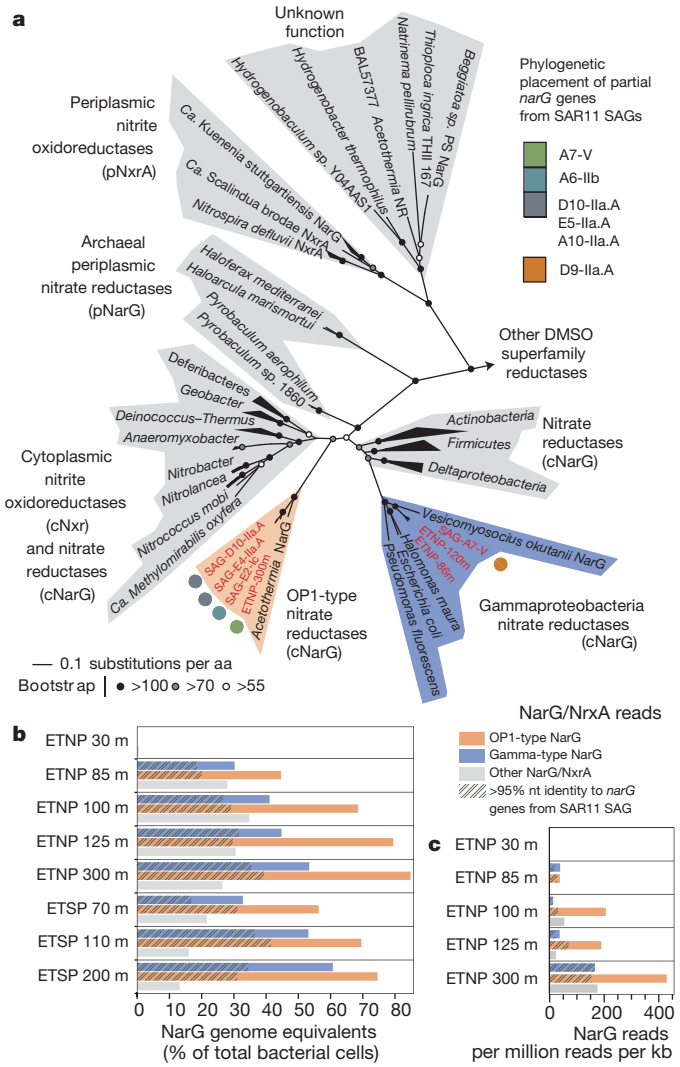


Figure 3 | Diversity, abundance and transcription of Nar enzymes in the OMZ. a, Phylogenetic reconstruction of NarG sequences identified in the SAR11 SAGs and metagenomic SAR11 contigs (ETNP prefix), along with reference Nar and Nxr enzymes. Partial gene sequences (represented with coloured pies) were subsequently added to the pre-constructed tree with phylogenetic placement. aa, amino acid. **b**, Relative abundance of NarG/NxrA enzymes in OMZ metagenomic data sets. Abundance was normalized to the *rpoB* gene abundance and thus represents genome equivalents, or the portion of OMZ bacterial cells that encode the enzyme. nt, nucleotide. **c**, Relative expression of NarG/NxrA proteins in the ETNP transcriptomes.

remote given the experimental and phylogenetic evidence, a positive relationship between NO_3^- reduction rates and the abundance of OP1 and Gamma-type genes and transcripts in the anoxic OMZ depths (Fig. 1c), and prior results showing O_2 sensitivity of OP1-type *nar* transcription[25] (Supplementary Discussion). Rather, the results strongly suggest that the identified SAR11 *narG* genes encode functional NO_3^- reductases.

SAR11 *nar* is abundant and highly transcribed

We next examined the abundance of SAR11-affiliated *nar* genes within the OMZ to evaluate the contribution of SAR11 cells to NO_3^- reduction. We first identified *nar* sequence reads in OMZ metagenomes using a similarity search-trained model that discriminates NO_3^- reductase (or NO_2^- oxidoreductase) reads from those of other genes of the DMSO superfamily (Supplementary Discussion). These *narG* reads were then classified within a reference phylogeny containing 320 NarG proteins, including OP1- and Gamma-type sequences. Remarkably, the majority of *narG* reads from OMZ metagenomes were classified as OP1- or Gamma-type Nar enzymes (Fig. 3b and Extended Data Fig. 6a), with the two variants accounting for 70% of total *narG* sequences at anoxic depths (Supplementary Table 4). Such high representation is consistent with quantitative (q)PCR-based counts of OP1- and Gamma-type *narG* copies at the collection site, where the two variants (summed) spiked at the OMZ NO_2^- maximum at >200,000 copies ml^{-1} (Extended Data Fig. 6b). The average number of *nar* genes per cell (that is, genome equivalents) was estimated by comparing the abundance of *nar* sequences with those of *rpoB*, a universal single-copy gene. On the basis of those estimations, Gamma and OP1 *nar* variants occur in up to 61% and 85% of OMZ bacteria, respectively (Fig. 3b and Extended Data Fig. 6b), assuming each *nar* type occurs once per genome. Such high values are striking but consistent with prior results taken from Basic Local Alignment Search Tool (BLAST)-based taxonomic assignments[18]. These values also exceed the estimated SAR11 abundances in the metagenomes, or those calculated directly from SAG 16S screening (up to 32% of the community), indicating that these gene variants occur in multiple copies per genome or in diverse bacteria (Supplementary Discussion). Metagenomic evidence suggests that the majority of these *nar* operons are found in SAR11 genomes within the OMZ. First, while our SAG collection captured only a fraction of total *nar* diversity, additional *nar* operons were identified in metagenomic contigs classified as SAR11 (Extended Data Figs 3, 7 and Supplementary Table 6). Second, the majority of the metagenomic *narG* reads showed >95% nucleotide identity with the *narG* genes encoded by the SAGs, suggesting that SAR11 cells are among the major contributors of Nar enzymes in the OMZ (Fig. 2b).

Metatranscriptome sequencing confirmed that SAR11-affiliated *nar* genes are transcribed in the OMZ. The abundance of both OP1- and Gamma-type variants in ETNP metatranscriptomes increased steadily from the lower oxycline (85 m) to the OMZ core (300 m), directly paralleling the abundance of the respective genes and the depth trend in NO_3^- reduction rates (Fig. 1c). Notably, within the ETNP OMZ, an average of 39% of all *narG* transcripts shared >95% nucleotide identity with the OP1- or Gamma-type sequences detected in SAR11 SAGs (Fig. 3c), a conservative lower-bound estimate of the contribution of SAR11 bacteria to the total *nar* transcripts within the OMZ. Accordingly, within the anoxic OMZ depths, *nar* genes are among the most transcriptionally active genes in the SAG genomes (Extended Data Fig. 8). The high transcriptional activity of SAR11 *nar* operons, interpreted alongside their distribution relative to NO_3^- reduction rates, suggests that SAR11 bacteria contribute substantially to community NO_3^- respiration.

Conclusions

Collectively, our findings identify diverse and abundant SAR11 lineages whose genome content and environmental distribution reflect adaptation to an anoxic niche, unlike all other SAR11 bacteria characterized to date. The experimentally verified NO_3^- reductase activity in the Gamma-type SAR11 *nar* variant, along with the high expression levels of divergent SAR11 *nar* genes in the functionally anoxic core of the OMZ, suggest that persistence in this niche is linked to NO_3^- respiration, consistent with the fundamental importance of this process in OMZs. Nitrate respiration in OMZs constitutes the primary mode for organic carbon mineralization and the main production route of NO_2^-, a critical substrate for the major nitrogen loss processes of anammox and denitrification. The presence and activity of *nar* operons in SAR11, as well as the high abundance of *nar*-associated SAR11 clades in the OMZ, implicate these versatile organisms as major contributors to the initiation of OMZ nitrogen loss. Together, these findings redefine the ecological niche of one of the planet's most dominant groups of organisms, providing a set of genomic references to establish SAR11 as a model for studies of nitrogen and carbon cycling in OMZs.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 16 December 2015; accepted 30 June 2016.

Published online 3 August 2016.

1. Brown, M. V., Schwalbach, M. S., Hewson, I. & Fuhrman, J. A. Coupling 16S-ITS rDNA clone libraries and automated ribosomal intergenic spacer analysis to show marine microbial diversity: development and application to a time series. *Environ. Microbiol.* **7,** 1466–1479 (2005).
2. Carlson, C. A. *et al.* Seasonal dynamics of SAR11 populations in the euphotic and mesopelagic zones of the northwestern Sargasso Sea. *ISME J.* **3,** 283–295 (2009).
3. Eiler, A., Hayakawa, D. H., Church, M. J., Karl, D. M. & Rappé, M. S. Dynamics of the SAR11 bacterioplankton lineage in relation to environmental conditions in the oligotrophic North Pacific subtropical gyre. *Environ. Microbiol.* **11,** 2291–2300 (2009).
4. Morris, R. M. *et al.* SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420,** 806–810 (2002).
5. Salter, I. *et al.* Seasonal dynamics of active SAR11 ecotypes in the oligotrophic Northwest Mediterranean Sea. *ISME J.* **9,** 347–360 (2015).
6. Thrash, J. C. *et al.* Single-cell enabled comparative genomics of a deep ocean SAR11 bathytype. *ISME J.* **8,** 1440–1451 (2014).
7. Giovannoni, S. J. *et al.* Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309,** 1242–1245 (2005).
8. Grote, J. *et al.* Streamlining and core genome conservation among highly divergent members of the SAR11 clade. *MBio* **3,** e00252–12 (2012).
9. Tripp, H. J. The unique metabolism of SAR11 aquatic bacteria. *J. Microbiol.* **51,** 147–153 (2013).
10. Konstantinidis, K. T., Braff, J., Karl, D. M. & DeLong, E. F. Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific subtropical gyre. *Appl. Environ. Microbiol.* **75,** 5345–5355 (2009).
11. Swan, B. K. *et al.* Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science* **333,** 1296–1300 (2011).
12. King, G. M., Smith, C. B., Tolar, B. & Hollibaugh, J. T. Analysis of composition and structure of coastal to mesopelagic bacterioplankton communities in the northern gulf of Mexico. *Front. Microbiol.* **3,** 438 (2013).
13. Vergin, K. L. *et al.* High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series Study site by phylogenetic placement of pyrosequences. *ISME J.* **7,** 1322–1332 (2013).
14. Paulmier, A. & Ruiz-Pino, D. Oxygen minimum zones (OMZs) in the modern ocean. *Prog. Oceanogr.* **80,** 113–128 (2009).
15. Tiano, L., Garcia-Robledo, E. & Revsbech, N. P. A new highly sensitive method to assess respiration rates and kinetics of natural planktonic communities by use of the switchable trace oxygen sensor and reduced oxygen concentrations. *PLoS One* **9,** e105399 (2014).
16. Kalvelage, T. *et al.* Nitrogen cycling driven by organic matter export in the South Pacific oxygen minimum zone. *Nature Geosci.* **6,** 228–234 (2013).
17. Stewart, F. J., Ulloa, O. & DeLong, E. F. Microbial metatranscriptomics in a permanent marine oxygen minimum zone. *Environ. Microbiol.* **14,** 23–40 (2012).
18. Ganesh, S., Parris, D. J., DeLong, E. F. & Stewart, F. J. Metagenomic analysis of size-fractionated picoplankton in a marine oxygen minimum zone. *ISME J.* **8,** 187–211 (2014).
19. Ganesh, S. *et al.* Size-fraction partitioning of community gene transcription and nitrogen metabolism in a marine oxygen minimum zone. *ISME J.* **9,** 2682–2696 (2015).
20. Ulloa, O., Canfield, D. E., DeLong, E. F., Letelier, R. M. & Stewart, F. J. Microbial oceanography of anoxic oxygen minimum zones. *Proc. Natl Acad. Sci. USA* **109,** 15996–16003 (2012).
21. Codispoti, L. A. *et al.* The oceanic fixed nitrogen and nitrous oxide budgets: Moving targets as we enter the anthropocene? *Sci. Mar.* **65,** 85–105 (2001).
22. Gruber, N. in *The Ocean Carbon Cycle and Climate* (eds Follows, M. & Oguz, T.) 97–148 (Springer, 2004).
23. Stewart, F. J., Sharma, A. K., Bryant, J. A., Eppley, J. M. & DeLong, E. F. Community transcriptomics reveals universal patterns of protein sequence conservation in natural microbial communities. *Genome Biol.* **12,** R26 (2011).
24. Lüke, C., Speth, D. R., Kox, M. A. R., Villanueva, L. & Jetten, M. S. M. Metagenomic analysis of nitrogen and methane cycling in the Arabian Sea oxygen minimum zone. *PeerJ* **4,** e1924 (2016).
25. Dalsgaard, T. *et al.* Oxygen at nanomolar levels reversibly suppresses process rates and gene expression in anammox and denitrification in the oxygen minimum zone off northern Chile. *MBio* **5,** e01966–14 (2014).
26. Kalvelage, T. *et al.* Oxygen sensitivity of anammox and coupled N-cycle processes in oxygen minimum zones. *PLoS ONE* **6,** e29299 (2011).
27. Dean, F. B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl Acad. Sci. USA* **99,** 5261–5266 (2002).
28. Dupont, C. L. *et al.* Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J.* **6,** 1186–1199 (2012).
29. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25,** 1043–1055 (2015).
30. Thrash, J. C. *et al.* Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. *Sci. Rep.* **1,** 13 (2011).
31. Luo, H. Evolutionary origin of a streamlined marine bacterioplankton lineage. *ISME J.* **9,** 1423–1433 (2015).
32. Rodríguez-Ezpeleta, N. & Embley, T. M. The SAR11 group of alpha-proteobacteria is not related to the origin of mitochondria. *PLoS ONE* **7,** e30520 (2012).
33. Viklund, J., Martijn, J., Ettema, T. J. G. & Andersson, S. G. E. Comparative and phylogenomic evidence that the alphaproteobacterium HIMB59 is not a member of the oceanic SAR11 clade. *PLoS ONE* **8,** e78858 (2013).
34. Konstantinidis, K. T. & DeLong, E. F. Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. *ISME J.* **2,** 1052–1065 (2008).
35. Takami, H. *et al.* A deeply branching thermophilic bacterium with an ancient acetyl-CoA pathway dominates a subsurface ecosystem. *PLoS ONE* **7,** e30559 (2012).
36. Kuwahara, H. *et al.* Reduced genome of the thioautotrophic intracellular symbiont in a deep-sea clam, *Calyptogena okutanii*. *Curr. Biol.* **17,** 881–886 (2007).
37. Iobbi, C., Santini, C.-L., Bonnefoy, V. & Giordano, G. Biochemical and immunological evidence for a second nitrate reductase in *Escherichia coli* K12. *Eur. J. Biochem.* **168,** 451–459 (1987).
38. Iobbi-Nivol, C., Santini, C. L., Blasco, F. & Giordano, G. Purification and further characterization of the second nitrate reductase of *Escherichia coli* K12. *Eur. J. Biochem.* **188,** 679–687 (1990).
39. Philippot, L. Denitrifying genes in bacterial and Archaeal genomes. *Biochim. Biophys. Acta* **1577,** 355–376 (2002).
40. Martinez-Espinosa, R. M. *et al.* Look on the positive side! The orientation, identification and bioenergetics of 'Archaeal' membrane-bound nitrate reductases. *FEMS Microbiol. Lett.* **276,** 129–139 (2007).
41. Rothery, R. A., Workun, G. J. & Weiner, J. H. The prokaryotic complex iron-sulfur molybdoenzyme family. *Biochim. Biophys. Acta* **1778,** 1897–1929 (2008).
42. Yoshimatsu, K., Iwasaki, T. & Fujiwara, T. Sequence and electron paramagnetic resonance analyses of nitrate reductase NarGH from a denitrifying halophilic euryarchaeote *Haloarcula marismortui*. *FEBS Lett.* **516,** 145–150 (2002).
43. Lücker, S. *et al.* A *Nitrospira* metagenome illuminates the physiology and evolution of globally important nitrite-oxidizing bacteria. *Proc. Natl Acad. Sci. USA* **107,** 13479–13484 (2010).
44. Starkenburg, S. R. *et al.* Genome sequence of the chemolithoautotrophic nitrite-oxidizing bacterium *Nitrobacter winogradskyi* Nb-255. *Appl. Environ. Microbiol.* **72,** 2050–2063 (2006).
45. Sorokin, D. Y. *et al.* Nitrification expanded: discovery, physiology and genomics of a nitrite-oxidizing bacterium from the phylum *Chloroflexi*. *ISME J.* **6,** 2245–2256 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the National Science Foundation (1151698 to F.J.S. and 1416673 to K.T.K.), the NASA Exobiology Program (NNX14AJ87G to J.B.G. and F.J.S.), the Sloan Foundation (RC944 to F.J.S.), and a Community Science Program grant from the Department of Energy (DOE; to F.J.S. and K.T.K.). The work conducted by the US DOE Joint Genome Institute, a DOE Office of Science User Facility, is supported under contract no. DE-AC02-05CH11231. L.A.B., M.L. and B.T. were supported by a European Research Council Advanced Grant (OXYGEN, 267233) and by the Danish National Research Foundation (DNRF53). D.T. acknowledges the support of Onassis Foundation Fellowship. We are grateful for the generosity of J. Cole, A. Magalon, C. Sohaskey and F. Sargent for providing *E. coli* mutant strains, S. Pavlostathis for the ion chromatography methods and J. Spain for his suggestions on the heterologous expression experiment.

Author Contributions D.T. conducted bioinformatics analyses. J.W., P.R. and N.S. conducted next-generation sequencing. C.C.P. and B.K.S. conducted qPCR analyses. L.M.R.-R developed additional bioinformatic methods for SAG contamination evaluation. R.R.M. and T.W. conducted cell sorting and SAG generation. L.A.B. and B.T. conducted process rate measurements. M.L. conducted STOX oxygen measurements. D.T., S.D., S.N., J.B.G. and A.S.B. conducted the heterologous expression experiments. F.J.S. and K.T.K. designed the study. F.J.S. and D.T. analysed the data and wrote the paper. All authors discussed the results and helped edit the manuscript.

Author Information SAR11 SAG sequences from the ETNP and GoM can be found in the BioProject database under accession numbers PRJNA290513 and PRJNA291283, respectively. The two OMZ metagenomes sequenced have been deposited in the Joint Genome Institute database under accession numbers 1059848 and 1059863. The mutant *E. coli* genome sequenced has been deposited in the BioProject database under accession number PRJNA322349. Sequences of the clone SAR11 *nar* operons have been deposited in GenBank under accession numbers KX275213 and KX275214. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.J.S. (frank.stewart@biology.gatech.edu).

Reviewer Information *Nature* thanks R. Kiene, D. Kirchman, R. Lasken and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Collection of ETNP and GoM samples for SAG analysis. No statistical methods were used to predetermine sample size. Selection of the SAR11 SAGs was randomized. The investigators were not blinded to allocation during experiments and outcome assessment. Samples were collected from the ETNP OMZ during the Oxygen Minimum Zone Microbial Biogeochemistry Expedition (OMZoMBiE) cruise (*R/V Horizon*, 13–28 June 2013). Sea water for single-cell sorting and single amplified genome (SAG) analysis was collected from two depths within the OMZ (125 m and 300 m) at station 6 (18° 54.0N, 104° 54.0W) on 19 June (Fig. 1). Additional ('control') samples were collected from a depth profile (1–2,107 m) of the Gulf of Mexico (GoM) on 29 May 2012 aboard the *R/V Endeavour* (cruise EN509) at station 5, with samples for SAG analysis preserved from the oxic surface (1 m). Collections were made using Niskin bottles on a rosette containing a conductivity–temperature–depth profiler (Sea-Bird SBE 911plus). Water samples were prepared by cryopreservation according to the protocol recommended by the Bigelow Single Cell Genomics Center. Briefly, triplicate 1 ml samples of bulk sea water (no prefiltration) were gently mixed with 100 µl of a glycerol TE stock solution (20 ml 100× TE pH 8.0, 60 ml sterile water, 100 ml glycerol) and frozen at −80 °C.

ETNP OMZ rate measurements, and oxygen and nutrient analysis. Samples for oxygen and nutrient measurements were collected on the same date and casts as those for single-cell sorting described above. Samples for rate measurements and metagenomics/transcriptomics (below) were collected a few hours later on the same day. Detailed collection and analysis procedures for those samples have been previously described[19]. Briefly, oxygen concentrations were determined using rosette-mounted sensors, including a SBE43 dissolved oxygen sensor for micromolar sensitivity and a high-resolution switchable trace amount oxygen (STOX) sensor for nanomolar-level measurements[46]. CTD-based oxygen measurements (SBE43) from three casts spanning this sampling period revealed no detectable movement in the oxycline, indicating stability in water column conditions.

Metagenome and metatranscriptome samples. Metadata, sequencing statistics, and accession numbers of all analysed metagenome and metatranscriptome data sets are in Supplementary Table 4. Here, we summarize the OMZ and GoM data sets at the core of our analysis. ETNP OMZ metatranscriptomic and metagenomic data sets were generated via MiSeq Illumina sequencing as described in ref. 19 and ref. 47, respectively, for 5 depths at station 6: the upper oxycline (30 m), lower oxycline (85 m), secondary chlorophyll maximum (100 m), secondary nitrite maximum OMZ (125 m) and OMZ core (300 m) (Supplementary Table 4). Metagenome data sets from the ETSP were generated by Roche 454 pyrosequencing as previously described[18] for 4 depths at an OMZ site (20° 05S, 70° 48W) off the coast of Iquique, Chile: the suboxic (<10 µM) upper OMZ just below the oxycline (70 m), the anoxic OMZ core (110 m, 200 m), and the oxic zone below the OMZ (1,000 m). The ETNP and ETSP data sets analysed here reflect the 0.2–1.6 µm biomass size fraction; this fraction was shown to contain the vast majority of bacterioplankton and SAR11 cells[19]. We also included two additional metagenomes, sampled on 5 May 2014 from the same site (station 6) in the ETNP, in order to obtain full-length *nar* operons for cloning purposes (see below). These metagenomes were obtained from depths of 68 m within the oxycline and 120 m within the OMZ. For the 9 GoM metagenomes released with this study, samples were collected from Niskin bottles (60 l per depth), and filtered on board using the same filtration systems as for the ETNP and ETSP metagenomes (0.2–1.6 µm fraction). DNA was extracted with the same protocol as for the OMZ samples[18] and libraries were prepared and sequenced in two lanes on an Illumina HiSeq (150 bp paired reads).

All metagenomic and metatranscriptomic data sets were quality trimmed as described below for the SAG data sets. The metatranscriptomic data sets were further filtered to remove rRNA transcripts using the SortMeRNA algorithm[48]. Four-hundred and fifty-four metagenomic data sets were filtered to remove duplicate sequences. The quality trimmed reads from the OMZ metagenomes (ETNP and ETSP), were assembled with IDBA[49] and genes were predicted on contigs longer than 500 bp with MetaGeneMark.hmm[50]. Taxonomic classification of metagenomic contigs was performed with MyTaxa[51]. *nar* operons were identified on metagenomic contigs as described below for the SAG assemblies.

SAG isolation and taxonomic characterization. Single amplified genomes (SAGs) were generated from individual bacterial cells[52], according to standard procedures in the Department of Energy Joint Genome Institute workflow[53]. Briefly, individual cells sorted on a BD Influx (BD Biosciences) were treated with Ready-Lyse lysozyme (Epicentre; 5 U/µl final concentration) for 15 min at room temperature before the addition of lysis solution. Whole-genome amplification was performed with the REPLI-g Single Cell Kit (Qiagen) in 2 µl reactions set up with an Echo acoustic liquid handler (Labcyte). Only the lysis and stop reagents from the REPLI-g kit received UV treatment since the amplification cocktail was pre-treated by the manufacturer. Amplification reactions were terminated after 6 h. PCR amplification and Sanger sequencing of a ~470 bp region of the 16S rRNA gene (amplified using primers 926wF (5′-AAACTYAAAKGAATTGRCGG-3′) and 1392R (5′-ACGGGCGGTGTGTRC-3′) for archaea and bacteria was used to assign a preliminary taxonomic identification to each of the SAGs, via comparisons to the Greengenes rRNA database.

SAG sequencing. A total of 27 SAR11 classified SAGs identified were randomly selected for sequencing, including 10 and 12 SAGs from 125 m and 300 m in the ETNP, respectively, and 5 'control' SAR11 SAGs from surface water (1 m) in the GoM. SAG DNA was prepared using the NexteraXT DNA Sample Prep kit (Illumina, San Diego, CA, USA) following the manufacturer's instructions. Libraries were pooled and sequenced at Georgia Tech on two runs of an Illumina MiSeq using a 500 cycle (paired end 250 × 250 bp) kit. Of the initial 27 SAGs, 8 were recovered in very low abundance in the read data or were removed due to potential contamination (>5%) as estimated with CheckM (see below) or the presence of 18S rRNA gene fragments, yielding the final set of 19 SAGs analysed here (Supplementary Table 1).

SAG sequence quality control assembly and functional gene annotation. Coupled reads were merged, when overlapping, using PEAR[54]. Both merged and un-merged reads were trimmed using SolexaQA++[55] with a PHRED score cutoff of 20 and a minimum fragment length of 50 bp. Illumina adaptors were clipped using Scythe (https://github.com/vsbuffalo/scythe) and reads were re-filtered for length (50 bp). Quality-trimmed reads were assembled with SPAdes[56]. Percentage of contamination and genome completeness were assessed based on recovery of lineage-specific marker gene sets using CheckM[29]. From the total of 27 SAG assemblies, 7 were excluded from the analysis due to low coverage (that is, less than 70 kb) or the presence of 18S rRNA sequences and BLASTP top matches to eukaryotic sequences reflecting contamination. For the remaining SAGs that passed the original quality control thresholds (Supplementary Table 1), when multiple fragments of a bacterial single-copy marker gene were identified, manual inspection of alignments revealed that multiplicity was due to assembly breaking points rather than contamination from divergent sequences, and such cases were retained for analysis (Supplementary Table 2). Evidence for contamination was detected in only one SAG, SAG A2 from the GoM, as multiplicity of divergent and nearly full-length marker genes. This SAG was excluded from further analysis.

For the final data set of 19 SAGs, coding sequences were predicted on scaffolds longer than 500 bp with GeneMark.hmm[50] and 16S rRNA gene sequences were identified using RNAmmer[57]. 16S rRNA sequences identified in the assemblies (4/4 GoM SAGs, and 8/15 OMZ SAGs) were compared to the 470 bp 16S fragment obtained during the initial SAG screening and confirmed to be identical. As an additional quality control step, all predicted genes from the 19 SAGs were taxonomically annotated using MyTaxa[51] and the taxonomic distributions of adjacent genes in the concatenated assembly (10 gene windows) were inspected for possible contamination. As discussed in Supplementary Discussion, a contaminant genome in the assembled contigs can be visualized in the MyTaxa scan plots (Extended Data Fig. 1).

Predicted genes were functionally annotated using the blast2go pipeline[58] for assignment to metabolic pathways, and screened manually for evidence of anaerobic energy metabolism. Detected genes of anaerobic metabolism, including nitrate reductase (*nar*) genes, as well as terminal oxidase genes and the single-copy marker gene *rpoB*, were further verified using HMMER3 (http://hmmer.janelia.org/) with default settings and recommended cutoffs for a match against available Pfam models[59]. Statistics of SAG quality control, assemblies, and contamination testing are in Supplementary Table 1 and 2.

Phylogenetic placement of SAGs. The evolutionary relatedness of SAR11 SAGs was assessed using the identified full or almost full-length 16S rRNA gene sequences from the assembled SAGs. For the SAGs from which no full-length 16S rRNA fragments were assembled, the shorter fragments obtained during screening were used in pairwise comparisons with full-length sequence references (Supplementary Table 3, 16S matrix). The 16S rRNA sequences from publicly available SAR11 genomes, as well as previously published 16S sequences[6,13] from subclades with no genome representatives, were included in the alignment to aid in the classification of the SAR11 subclades. Additionally, genome representatives of divergent alphaproteobacteria classes, as well as a beta- and gammaproteobacterium were included to facilitate the rooting of the tree. Maximum likelihood phylogenetic reconstruction was performed with RAxML with 1,000 bootstraps and the GTR model for nucleotides[60]. Additionally, hidden Markov models (HMMs) of 106 housekeeping genes found in single copy in bacterial genomes were used to identify marker genes in available SAGs and reference genomes using HMMER3 (http://hmmer.janelia.org/) with default settings and the recommended cutoff[28]. The identified marker genes (Supplementary Table 1) were aligned using Clustal Omega[61] and the protein alignments concatenated using Aln.cat.rb from the enve-omics collection (http://enve-omics.ce.gatech.edu/) to remove invariable sites and maintain protein coordinates. The concatenated alignment was used to build a maximum likelihood phylogeny with RAxML, using 1,000 bootstraps, and the PROTGAMMAAUTO function, which identifies the best amino acid substitution model for each protein. SAGs where assigned to SAR11 subclades based

on the consensus categorization of both 16S rRNA and marker gene phylogenies, in accordance with previously published subclade identification sequences[6,13]. OMZ-derived SAR11 SAGs from the SAR11 IIa lineage were further categorized as subclade IIa.A, to differentiate them from the currently available reference SAR11 IIa representative (HIMB058), classified here as subclade IIa.B. Average amino acid identities (AAIs) were estimated as described previously[62].

***nar* functional gene validation and phylogeny.** Reference nitrate reductase and nitrite oxidoreductase protein sequences ($n = 697$) representing divergent bacterial and archaeal phyla were downloaded from UniProt/Swiss-Prot[63], together with representatives of other DMSO family oxidoreductases ($n = 71$), using as a guide the reference tree from ref. 64. From this 697-sequence set, 321 full-length NarG/NxrA sequences were selected to represent all the clades, along with the 71 additional non-NarG/NxrA proteins. The NarG/NxrA subset included the closest relatives to the SAG OP1 and Gamma-type Nar variants, as determined by BLAST. All protein sequences ($n = 392$), including the full-length NarG identified in the SAGs, were aligned with Clustal Omega, and a maximum likelihood phylogeny was reconstructed with RAxML with 1,000 bootstraps and the PROTGAMMAAUTO model. Partial fragments of the NarG protein were then added to the alignment using MAFFT's 'addfragments'[65], and the evolutionary placement algorithm (EPA) implemented in RAxML was used to place them within the reference tree[66]. The same procedure was followed for the phylogenetic reconstruction and placement of identified NarH protein sequences.

Quantification of *narG*-encoding reads from the metagenomes and metatranscriptomes was done using BLAST searches against a manually curated NarG database and the software ROCker (L. H. Orellana, L. M. Rodriguez-R and K. T. Konstantinidis, manuscript submitted). Using receiver-operator curve (ROC) analysis, ROCker identifies the most discriminant BLAST bit-score per position in a reference alignment (NarG database) given a certain read length by simulating *in silico* metagenomic data sets that include the reference genes. This strategy permits the accurate estimation of abundance of target genes in short-read data sets, minimizing false negatives and positives derived from closely related proteins or conserved domains, a critical challenge in the detection of *narG* due to the ubiquity of other closely related DMSO oxidoreductases. The NarG database was manually curated and confirmed by the phylogenetic reconstruction of all available nitrate reductase and nitrite oxidoreductase sequences and visual inspection of the multi-sequence alignment for conservation of known functional domains and motifs. The final NarG database consisted of 697 nitrate reductases/nitrite oxidoreductases (positive set) and 71 representative non-NarG/NxrA DMSO family proteins (negative set for identification of false positive BLAST matches). All data sets, as well as the ROCker models built for *narG* quantifications in metagenomes with different read lengths, are available at http://enve-omics.ce.gatech.edu/rocker/. Additionally, the model for the identification of *rpoB* fragments in metagenomes was used to estimate coverage of *rpoB* in metagenomes.

The abundance of *narG* sequences in meta-omic data sets was estimated as genomic equivalents for each sample, by normalizing the coverage of *narG* for the gene length (reads per nucleotide of *narG*), and dividing the normalized value by the *rpoB*-normalized coverage (reads per nucleotide of *rpoB*) as shown in Supplementary Table 4. To quantify the abundance of the *narG* variants (OP1, Gamma-type), protein fragments were predicted in all identified (from ROCker) *narG* reads using FragGeneScan[67] and placed in the reference DMSO tree using RAxML-EPA. The abundances of the OP1-type or Gamma-type variants were estimated based on the number of reads that were placed in the terminal or internal nodes of the aforementioned clades on the reference tree, using JPlace.to_iToL.rb from the enve-omics collection. The NarG metagenomic reads (predicted open reading frames (ORFs)) placed within those nodes, were used to construct the recruitment plots shown in Extended Data Fig. 8b. BLASTP was used to map the reads against the reference NarG sequences, and the recruitment plots were constructed with the BlastTab.catsbj.pl and BlastTab.recplot.R scripts from the enve-omics collection.

Thus, the reported abundances of OP1 and Gamma-type *narG* in metagenomes/metatranscriptomes are based on phylogenetic assignment of *nar* reads, rather than a strict sequence similarity cutoff. To estimate a lower limit for the abundance of NarG sequences presumably encoded by SAR11 genomes, the number of reads with more than 95% nucleotide identity to the reference NarG sequences found in the SAGs was estimated, and shown in Extended Data Fig. 6b, c. The figure shows abundance estimates for reads that are phylogenetically assigned to OP1 and Gamma nodes, with partitioning of the data into reads that share less than and greater than 95% nucleotide identity with the SAG OP1 and Gamma-type references.

NarG divergence in reference closed genomes. Identification of NarG in all closed genomes available from GOLD (27,461 bacterial and 685 archaeal genomes)[68] was performed using HMMER3 with default settings. The results were further refined by a competitive BLAST search[69] against the custom-made NarG reference database (used for ROCker), which included DMSO family oxidoreductase enzyme reference sequences. Matches with best hit against NarG sequences and a bit score higher than 900 were annotated as nitrate reductases or nitrite oxidoreductases. When found in multiple copies (up to 6), a reciprocal BLASTP search was performed to estimate sequence divergence, measured as amino acid identity.

Quantification of SAR11 clades in metagenomes and metatranscriptomes. For each metagenome/metatranscriptome, reads potentially derived from SAR11 genomes were identified by a competitive BLAST best-match approach. A custom database was built using all available closed genomes from NCBI-ftp (2638 bacterial, 165 archaeal) and 39 genome representatives of the SAR11 lineage, including 20 published isolate or SAG sequences and the 19 SAG sequences produced in this study (Supplementary Table 1). Metagenomic and metatranscriptomic reads (predicted ORFs with FragGeneScan) were then compared against the database using BLASTP, and the subset of reads with a best match against any of the SAR11 genomes and an e value < 0.001 was classified as 'SAR11 reads' (Supplementary Table 4). To quantify the relative abundance of distinct SAR11 subclades, the SAR11 reads were further classified as follows. We used the coverage of marker genes that could be found in all the subclades to more accurately estimate the abundance of distinct subclades and overcome both the biased representation of SAR11 subclades in the available genomes, and the partial nature of SAG genomes. For all 39 available SAR11 genomes, 5,707 orthologous genes (OGs) were identified by reciprocal best match and Markov clustering with inflation 1.5 using ogs.mcl.rb from the enve-omics collection. From the identified OGs, 507 were represented at least once in each of the 8 subclades. All metagenomic and metatranscriptomic reads (SAR11 subsets) were mapped against the database containing all protein sequences from the 507 OGs (which were tagged according to subclade of origin) using the BLASTX option from Diamond[70] and only the best matches for each read were kept. The coverage of each OG for each subclade was estimated based on that competitive best match result, normalized for the gene length (reads per bp of each OG), and the average coverage of all 507 OGs was used to estimate the abundance of subclades. Additionally, the number of *rpoB* reads for each metagenome was identified (for either the total data set or the subset of the SAR11 reads), and the coverage of *rpoB* was used as a normalization factor to estimate the abundance of SAR11 subclades over the total bacterial community.

Functional characterization of SAR11 *nar* operons. A previously constructed NO_3^- reductase deficient *Escherichia coli* strain[71] was used as the genetic system for heterologous expression of SAR11 *nar* genes. We used whole-genome sequencing (Illumina MiSeq) to confirm that this strain lacked all three NO_3^- reductases (Δ*narGI* Δ*napAB narZ*::Ω; Extended Data Fig. 5). The phenotype of this strain, hereafter referred to as the triple mutant, was verified by a lack of NO_2^- production and an absence of growth with NO_3^- under anaerobic conditions, compared to the wild-type MC4100 *E. coli* strain (Extended Data Fig. 5).

Complete sequences from one OP1-type, and one Gamma-type *nar* operon, containing upstream and downstream sequences, were identified from the ETNP 300 m and ETNP 120 m metagenomes (see above). These sequences were confirmed to be identical to the operons in SAG A7 (which was lacking part of the N terminus of the *narG* gene; Extended Data Fig. 3). Purified DNA from the ETNP 300 m and ETNP 120 m metagenomic samples was used as template for PCR amplification. In addition, we used genomic DNA from *E. coli* strain K12 MG1655 as a positive control. Because metagenomic samples are usually fragmented and the entire *nar* operon is 6.9 kb, primers were designed to amplify the OP1-type, Gamma-type and *E. coli* wild-type operon in two blocks. The first block spanned from the native NarG ribosome binding site to the end of the *narG* gene, and the second block included the end of the *narG* gene to the *narI* stop codon. The resulting PCR products were gel purified, assembled and cloned into pBbA1K, a low-copy vector including the IPTG-inducible pTrc promoter[72] by In-Fusion cloning (Clontech, Mountain View, CA). The cloning reactions were transformed into TOP10 cells, and inserts were sequence-verified by Pacbio sequencing (Pacific Biosciences, Menlo Park, CA). The final *nar* sequences were identical (OP1 operon, and NarG,I proteins of Gamma operon) or nearly identical with silent substitutions (99% and 98% AAI for the Gamma-type NarG and H proteins) compared to the sequences from SAG A7 (GenBank accessions KX275213, KX275214). Correct clones were isolated for each operon type, and purified plasmid was used to electroporate the triple mutant *E. coli* strains described above to generate recombinant strains expressing the heterologous *nar* operons for functional characterization.

For anaerobic cultures performing NO_3^- respiration, strains were first induced in LB medium with 0.5 mM IPTG for 5 h, and 20 μl of inoculum was subsequently introduced in gas tight tubes under N_2 atmosphere. The medium was prepared as previously described[73], composed from potassium phosphate buffer (100 mM, pH 7.4), 15 mM $(NH_4)_2SO_4$, 9 mM NaCl, 2 mM $MgSO_4$, 5 μM Na_2MoO_4, 10 μM Mohr's salt, 100 μM $CaCl_2$, 0.5% casaminoacids and 0.01% thiamine. Glycerol (40 mM) was used as the sole carbon source, and NO_3^- was added at 30 mM. IPTG (0.5 mM), kanamycin (30 μg/ml) and streptomycin (30 μg/ml) were used with the

recombinant strains. Samples for NO_3^- and NO_2^- concentrations were obtained at regular time intervals during incubations, filtered through 0.2 μm porosity filters and injected into a Dionex DX ion chromatography unit with the Dionex IonPac AS14A analytical column[74]. Growth in incubations was assessed as optical density ($OD_{600\ nm}$). Growth curve data from replicated cultures (triplicate) were fitted to a logistic model with variables r (specific growth rate), P_0 (initial population) and K (carrying capacity), using nonlinear least-squares estimates and prediction of OD per time point with confidence intervals as implemented in enve.growthcurve from the enve-omics collection (http://enve-omics.ce.gatech.edu/).

Nitrate reductase activity was further verified in cell lysates from cells grown anaerobically for 12 days. Cells resuspended in 100 mM sodium phosphate buffer (pH 7.2) containing 0.02% Tween 80 were lysed by sonication in a Bioruptor UCD-200 (Diagenode). Protein concentration of the cell lysate was determined using a Qubit 2.0 fluorometer (Thermo Fisher Scientific) and 100 μg of protein was added to a reaction containing 100 mM $NaNO_3$ and benzyl viologen as electron donor. The reaction was bubbled with N_2 for 2 min before initiation with the addition of 50 μl of 30 mM sodium dithionite in 10 mM NaOH (final volume: 500 μl). Aliquots (50 μl) were removed at 20 min intervals and NO_2^- concentration determined colorimetrically after the addition of 50 μl Griess reagent (prepared with equal volumes of 0.1% *N*-1-napthylethylenediamine dihydrochloride in water and 1% sulfanilamide in 5% phosphoric acid). All assays were performed in triplicate. Finally, NO_2^- production from NO_3 was further confirmed using whole cell assays with of 8 replicate clones (per recombinant strain) grown aerobically on 96-well plates in 70 μl Luria–Bertani (LB) broth supplemented with 30 mM NO_3^- and various IPTG concentrations. Nitrite production was identified via the Griess reaction as described above.

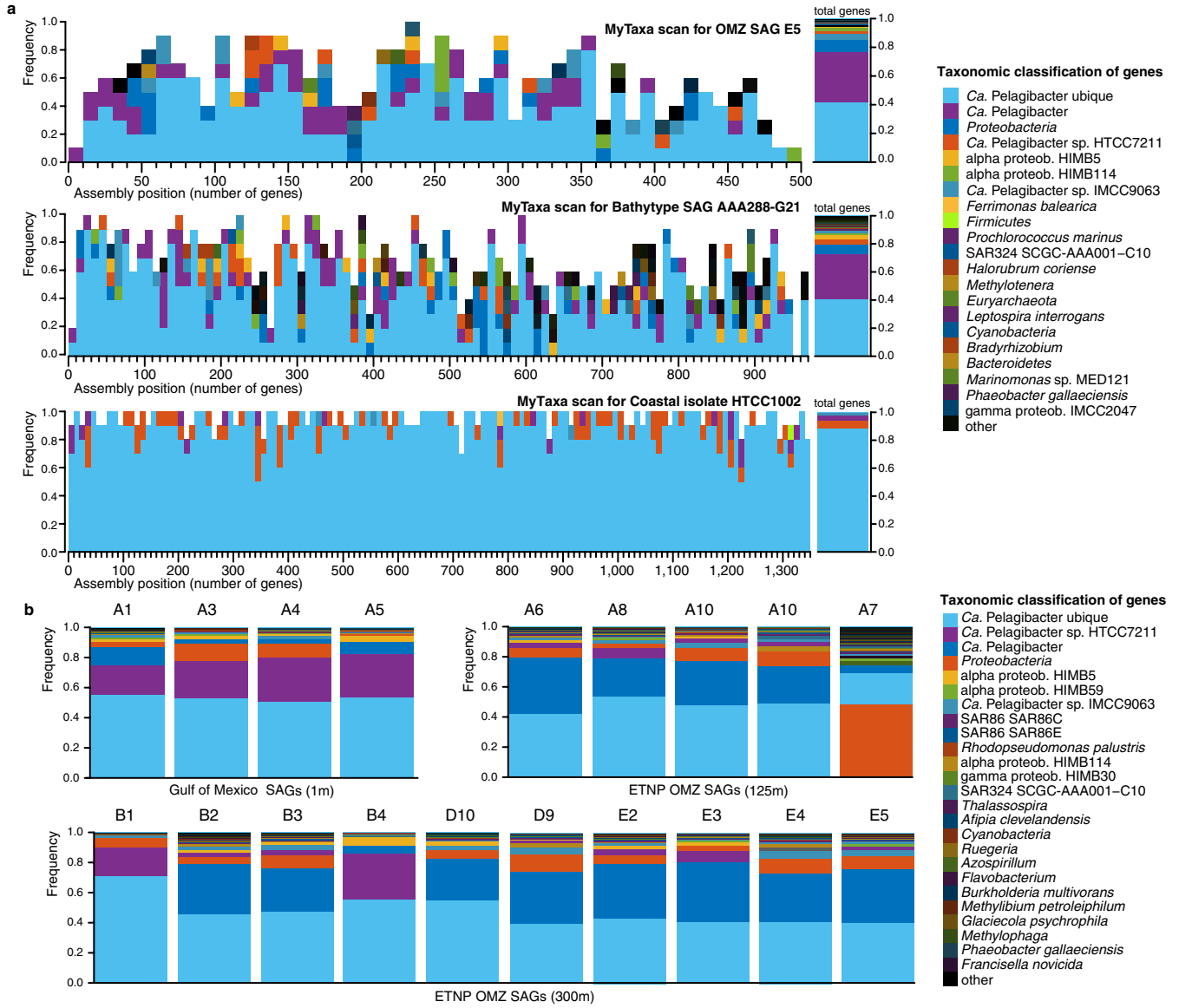
Quantitative PCR of 16S rRNA genes and SAR11 *nar* variants. Quantitative PCR (qPCR) was used to count OP1- and Gamma-type *narG* and total bacterial 16S rRNA gene copies. Sea-water samples for qPCR were collected in 2014 from three sites in the ETNP, including station 6 from which the SAG samples were obtained.

Primers for *narG* PCR were designed based on alignments of *narG* sequences recovered from OMZ SAGs, targeting sites inclusive of all OMZ SAR11-affiliated *nar* variants and exclusive of *narG* from the closest database reference sequences. Primer selection resulted in the following: GammaF, 5′-GCGTAAAATAATTT CTTCTCCTACATGGA-3′; and GammaR, 5′-AGTTCAATCCAGTCATTAT CTTCTACATC-3′ amplifying a 401-nucleotide fragment of the Gamma-type *nar*; and OP1F, 5′-ACCATCAAGGAATAAGAGAATTAGG-3′; and OP1R, 5′-TGGATTCCGTTTTCACAATACATTTC-3′ amplifying a 288-nucleotide fragment of the OP1-type *nar*. PCR reactions were performed with DNA template from the OMZ 300 m sample (station 6) and the oxic Gulf of Mexico as a negative control with the following conditions: incubation at 50 °C for 2 min, 95 °C for 10 min, followed by 40 cycles of denaturation at 95 °C (15 s) and annealing at 53 °C (for OP1) and 54 °C (for Gamma) (1 min each). Amplicons with the expected length were observed only in the OMZ sample and were purified and concentrated using the QIAquick PCR purification kit (Qiagen). Clone libraries were prepared with the TOPO TA cloning kit (Life Technologies) following the manufacturer's protocol, and plasmids from overnight grown selected colonies were isolated with the PureLink Quick Plasmid Miniprep Kit (Life Technologies). Inserts were purified using the QIAquick PCR Purification kit and sequenced on an Applied Biosystems 3730xl DNA Analyzer using BigDye Terminator v.3.1 cycle Sanger sequencing (Life Technologies). Sequencing recovered 14 sequences generated using OP1 primers and 12 generated using Gamma primers. All OP1-like sequences were most closely related (via BLASTX against the NCBI-nr database) to *narG* of an uncultured *Acetothermia* bacterium OP1 (dbj|BAL57372.1|), whereas all Gamma-like sequences were most closely related to the gammaproteobacterial endosymbiont of *Calyptogena okutanii* (*Ca.* Vesicomyosocius okutanii; ref|WP_011930032.1|), consistent with the phylogenetic classification of the recovered SAG *nar* sequences as described in the main text and confirming the specificity of the primer sets. However, sequences within each clone set shared on average 96% (OP1 set) and 93% (Gamma set) nucleotide identity, raising the possibility that our primer sets may not amplify all OP1 and Gamma-type *nar* variants in the community. We therefore consider our abundance estimates to be lower bounds.

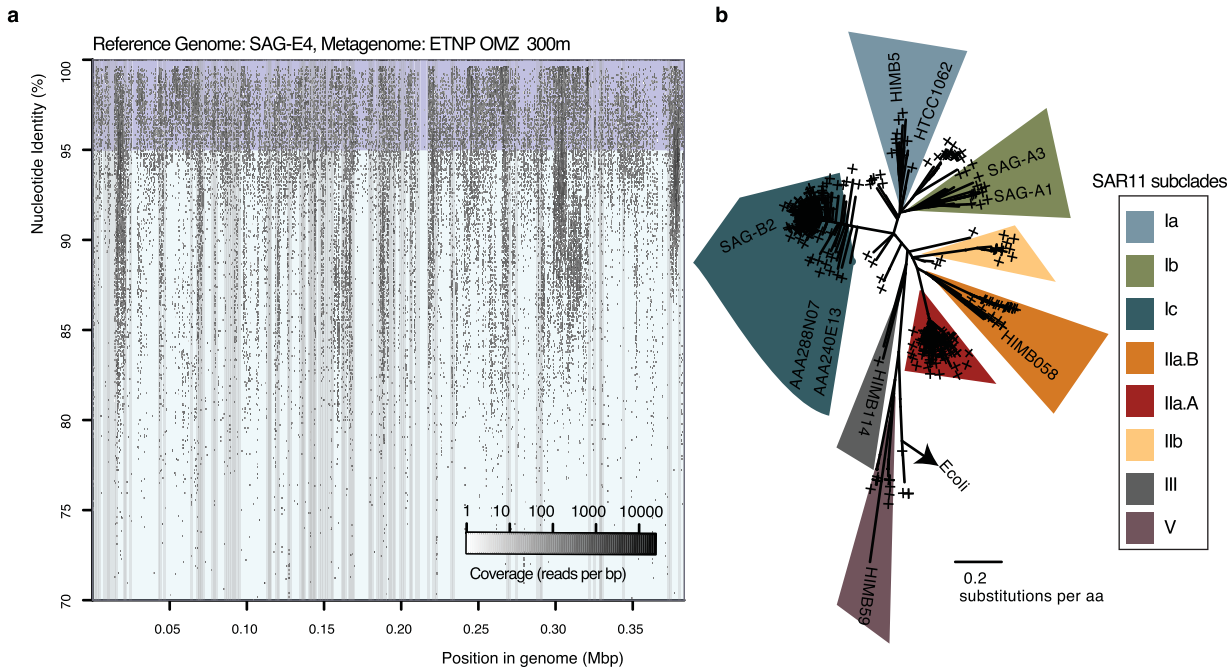
The OP1 and Gamma primer sets, along with universal bacterial 16S rRNA gene primers 1055f and 1392r, were used for SYBR Green-based qPCR. Tenfold serial dilutions of DNA from a plasmid carrying *narG* amplicons (described above) and a single copy of the 16S rRNA gene (from *Dehalococcoides mccartyi*) were included on each qPCR plate and used to generate standard curves, with a detection limit of ~30 and 10–15 gene copies/ml for 16S rRNA and *narG* variants, respectively. Assays were run on a 7500 Fast PCR System and a StepOnePlus Real-Time PCR System (Applied Biosystems). All samples were run in triplicate with conditions as follows: 2 min incubation at 50 °C, followed by 10 min at 95 °C followed by 40 cycles of denaturation at 95 °C (15 s) and annealing at 60 °C (1 min).

Data availability. Metagenomes from the GoM are available from the BioProject database under accession number PRJNA291283. Sample accession numbers and further information on all metagenomics and SAG data sets used in this study are provided in Supplementary Tables 4 and 1, respectively.

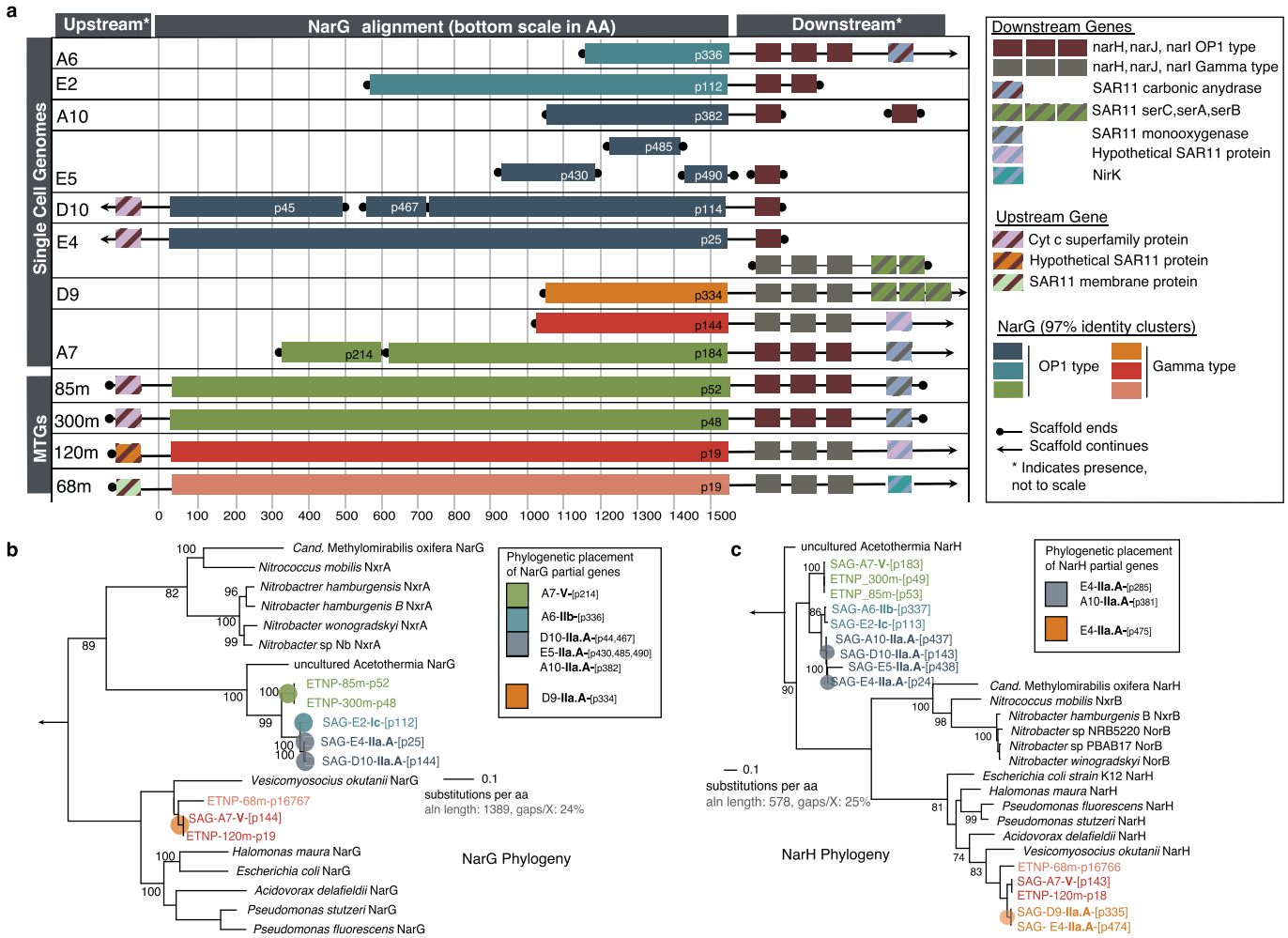
46. Revsbech, N. P. *et al.* Determination of ultra-low oxygen concentrations in oxygen minimum zones by the STOX sensor: STOX oxygen sensor. *Limnol. Oceanogr. Methods* **7,** 371–381 (2009).
47. Glass, J. B. *et al.* Meta-omic signatures of microbial metal and nitrogen cycling in marine oxygen minimum zones. *Front. Microbiol.* **6,** 998 (2015).
48. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28,** 3211–3217 (2012).
49. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28,** 1420–1428 (2012).
50. Zhu, W., Lomsadze, A. & Borodovsky, M. *Ab initio* gene identification in metagenomic sequences. *Nucleic Acids Res.* **38,** e132 (2010).
51. Luo, C., Rodriguez-R, L. M. & Konstantinidis, K. T. MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Res.* **42,** e73 (2014).
52. Raghunathan, A. *et al.* Genomic DNA amplification from a single bacterium. *Appl. Environ. Microbiol.* **71,** 3342–3347 (2005).
53. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499,** 431–437 (2013).
54. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30,** 614–620 (2014).
55. Cox, M. P., Peterson, D. A. & Biggs, P. J. SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* **11,** 485 (2010).
56. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19,** 455–477 (2012).
57. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35,** 3100–3108 (2007).
58. Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21,** 3674–3676 (2005).
59. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42,** D222–D230 (2014).
60. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).
61. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7,** 539 (2011).
62. Konstantinidis, K. T. & Tiedje, J. M. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr. Opin. Microbiol.* **10,** 504–509 (2007).
63. Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B. & Wu, C. H. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31,** 926–932 (2015).
64. Castelle, C. J. *et al.* Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. *Nature Commun.* **4,** 2120 (2013).
65. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).
66. Berger, S. A., Krompass, D. & Stamatakis, A. Performance, accuracy, and Web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst. Biol.* **60,** 291–302 (2011).
67. Rho, M., Tang, H. & Ye, Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* **38,** e191 (2010).
68. Reddy, T. B. K. *et al.* The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Res.* **43,** D1099–D1106 (2015).
69. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
70. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12,** 59–60 (2015).
71. Potter, L. C., Millington, P., Griffiths, L., Thomas, G. H. & Cole, J. A. Competition between *Escherichia coli* strains expressing either a periplasmic or a membrane-bound nitrate reductase: does Nap confer a selective advantage during nitrate-limited growth? *Biochem. J.* **344,** 77–84 (1999).
72. Khlebnikov, A. & Keasling, J. D. Effect of *lacY* expression on homogeneity of induction from the P_{tac} and P_{trc} promoters by natural and synthetic inducers. *Biotechnol. Prog.* **18,** 672–674 (2002).
73. Alberge, F. *et al.* Dynamic subcellular localization of a respiratory complex controls bacterial respiration. *eLife* **4,** e05357 (2015).
74. Hajaya, M. G. & Pavlostathis, S. G. Fate and effect of benzalkonium chlorides in a continuous-flow biological nitrogen removal system treating poultry processing wastewater. *Bioresour. Technol.* **118,** 73–81 (2012).
75. Bender, K. S. *et al.* Identification, characterization, and classification of genes encoding perchlorate reductase. *J. Bacteriol.* **187,** 5090–5096 (2005).
76. Jormakka, M., Richardson, D., Byrne, B. & Iwata, S. Architecture of NarGH reveals a structural classification of Mo-bisMGD enzymes. *Structure* **12,** 95–104 (2004).



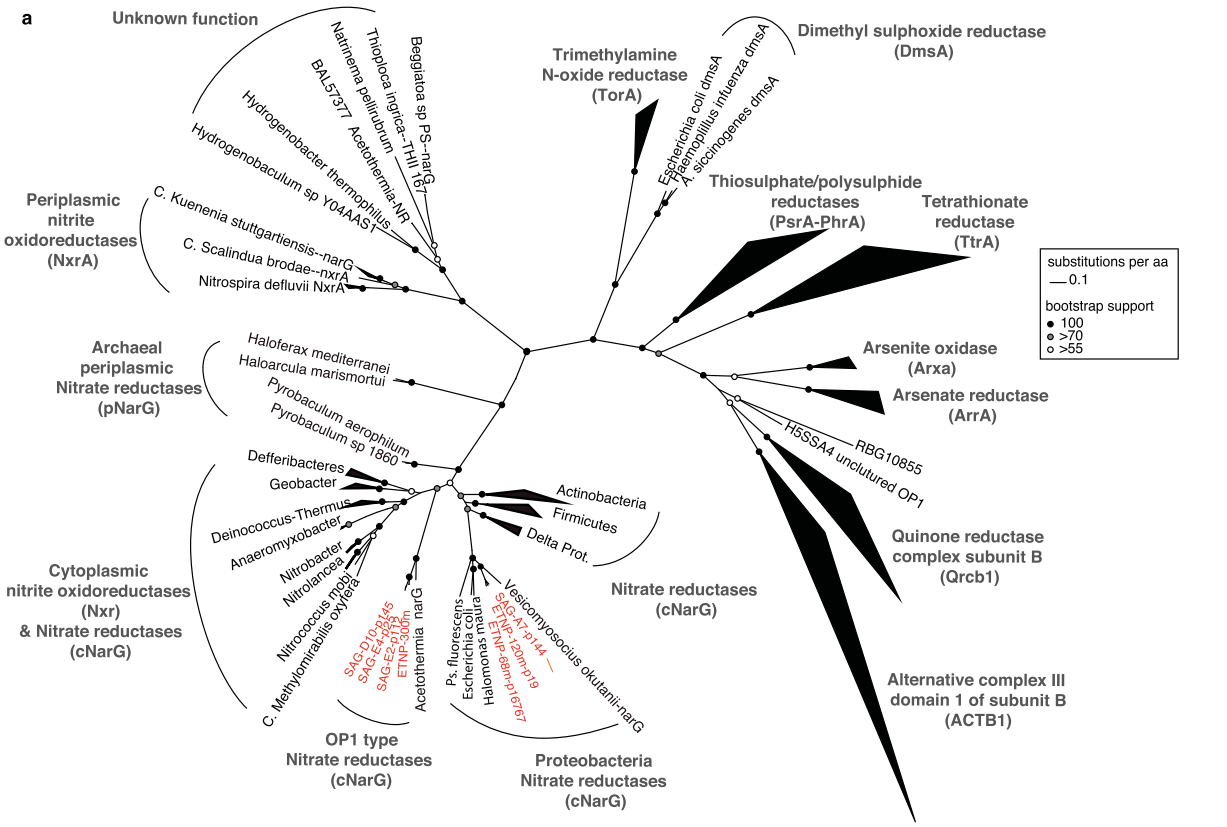
Extended Data Figure 1 | Evaluation of contamination based on MyTaxa taxonomic affiliations. a, Representative MyTaxa plots to test for contamination based on taxonomic affiliations of predicted genes. The MyTaxa algorithm[51] predicts the taxonomic affiliation on the basis of a weighted classification scheme that takes into account the phylogenetic signal of each protein family. Each gene is assigned to the deepest taxonomic resolution (out of phylum, genus and species) for which a high-confidence value can be obtained (score 0.5). Each MyTaxa scan represents taxonomic distributions of all the predicted genes for one genome, given in windows of 10 genes, and sorted based on their position in the concatenated assembly of the genome (when a partial genome is used). **a**, **b**, White space in the histograms represents genes that could not be assigned to a given taxon due to (1) lack of BLASTP hits against the reference database (a collection of closed and draft genomes) or (2) lack of high confidence scores. Notice that for the representative OMZ SAG E5, more than 80% of the genes can be classified as *Candidatus* Pelagibacter (SAR11), with an additional 10% assigned to *Proteobacteria*. Note there are no genome representatives for this taxon (that is, SAR11 subclade IIa.A) in the database upon which MyTaxa is based. Similar results are obtained for the bathytype SAR11 SAG[6], as this genome also lacks representatives. The closed genome from a coastal isolate HTCC1002 is shown for comparison to demonstrate a typical pattern for cases when close relatives of the query genome are available in the reference database, as is the case for this isolate. **b**, Taxonomic classifications of genes from the 19 SAGs analysed here. Each distribution was obtained from the MyTaxa scans performed for each SAG. The percentage of the total genes that could be taxonomically classified with MyTaxa was on average ~60%, and varied depending on the completeness of the genome (that is, partial genes are less likely to be assigned taxonomy with high confidence). These values are also reported in Supplementary Table 1. Of the genes that could be classified, the majority (>90%) were classified to SAR11 taxa.



Extended Data Figure 2 | Microdiversity within the SAR11 populations. **a**, Recruitment plot of metagenomic reads from the ETNP OMZ 300 m sample, against scaffolds from SAG E4. Notice that the recruited reads vary in identities from 100% down to 85%, indicating the presence of closely affiliated clades, as well as extensive microdiversity within the same clade (that is, reads sharing >95% identity). **b**, Phylogenetic reconstruction of reference RpoB protein sequences from SAR11 genomes, and placement of identified RpoB metagenomic sequences (denoted with the cross symbols). The alignment length was 1,406 columns with 5.9% gaps or undetermined sites. The presence of multiple divergent *rpoB* reads within the same subclade (predominantly for subclades IIa.A and Ic) suggests high abundance but also extensive microdiversity within those populations (rather than clonal populations).

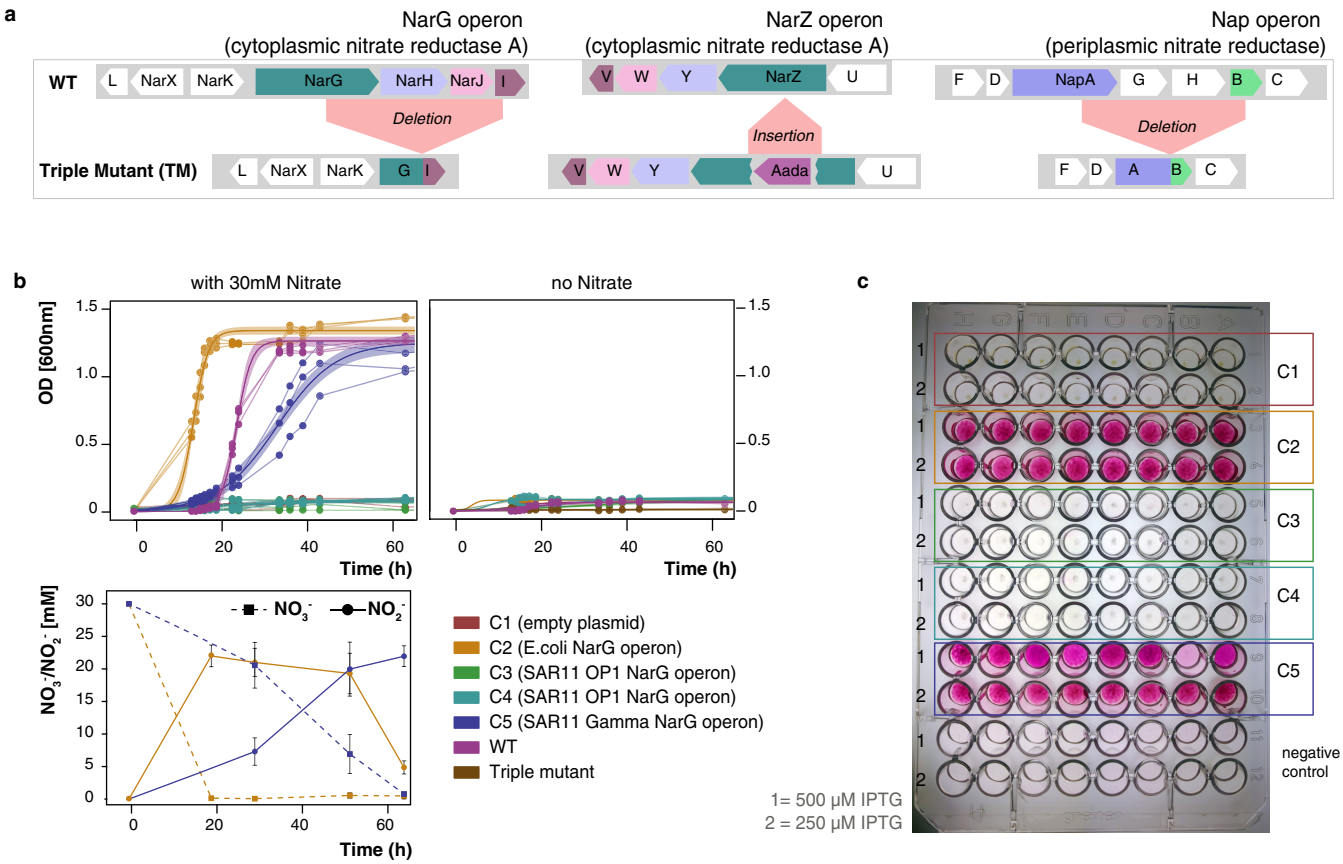


Extended Data Figure 3 | *nar* genes encoded by SAR11 populations of OMZs. a, *nar* operon and adjacent genes identified in SAR11 SAGs from the ETNP OMZ, and in assemblies from the 85 m and 300 m ETNP OMZ metagenomes. *narG* sequences with at least 97% amino acid similarity are represented with the same colour. **b**, **c**, Representative maximum likelihood phylogeny to show sequence variation among full-length or near full-length *narG* (**b**) and *narH* (**c**) amino acid sequences identified in the SAGs. A subset of cytoplasm-oriented Nar and Nxr enzymes from publicly available genomes is also included. A comprehensive phylogeny showing the placement of SAR11 *nar* sequences relative to enzymes ($n = 392$) of the DMSO family is in Fig. 2a. Coloured pies represent the placement of shorter *narG*/*narH* gene fragments identified in the SAGs. Bootstrap values over 50 are shown. Outgroups (arrows) are *E. coli dmsA* (**b**) and *dmsB* (**c**). Note that the Gamma-type *nar*-containing contig recovered in E4 (Fig. 2a) contains *narHJI*, but not *narG*; E4 Gamma-type is therefore not represented in Fig. 3b. All genes co-localized in the *nar*-containing contigs are listed in Supplementary Table 5. The p-numbers are gene identifiers given by the gene prediction software, consistent with those in Supplementary Table 5.

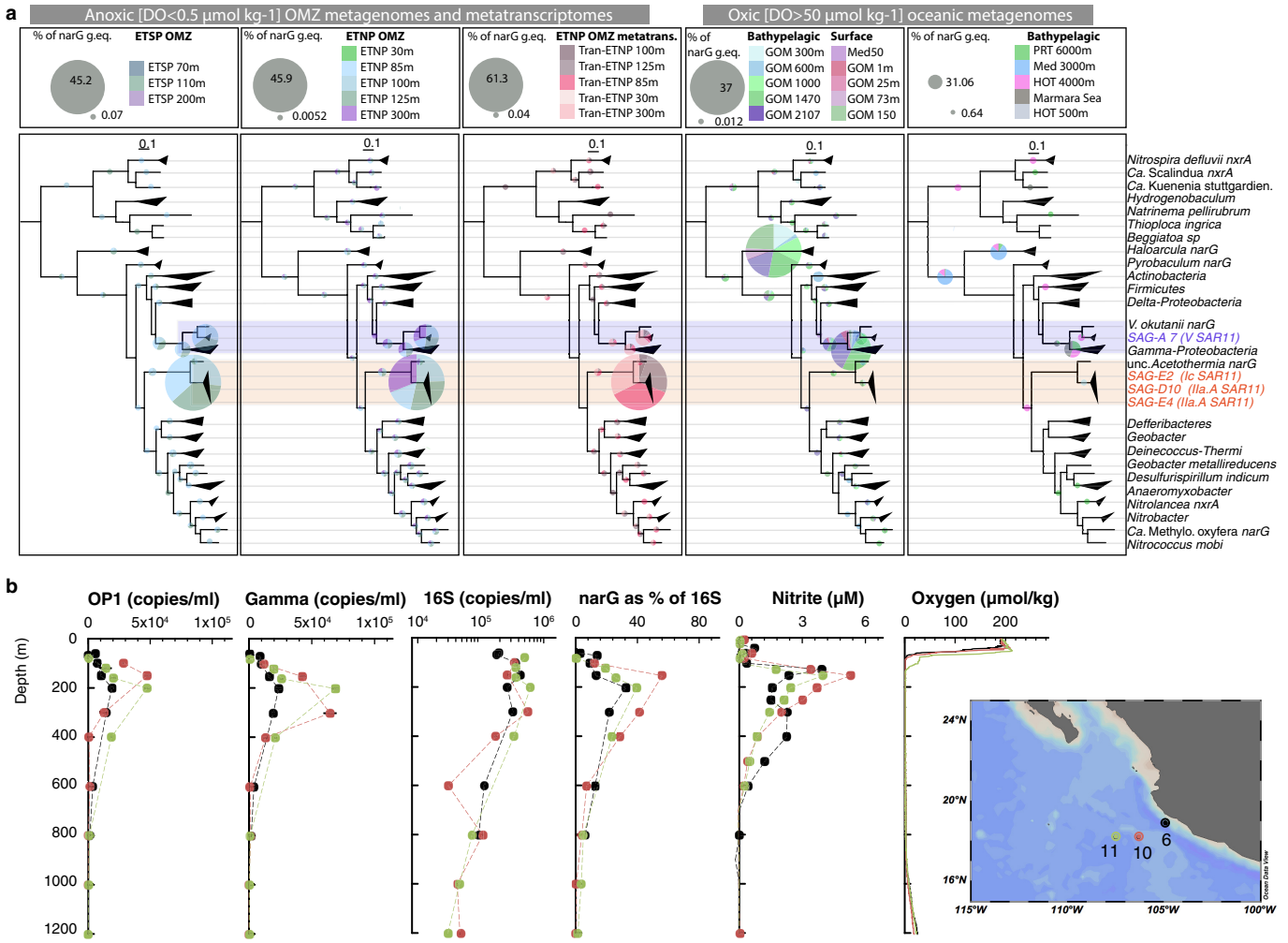


Extended Data Figure 4 | See next page for caption.

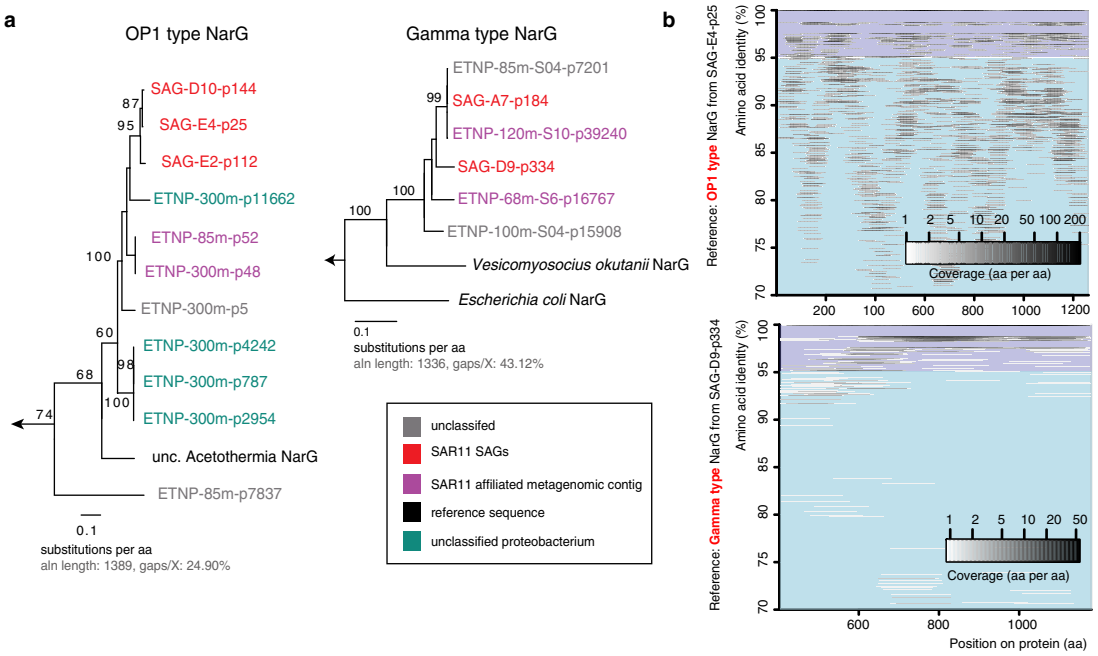
Extended Data Figure 4 | Identified NarG in SAR11 SAGs are members of the DMSO superfamily of oxidoreductases. a, Phylogenetic reconstruction of NarG and DMSO enzymes. The tree shown in Fig. 2 is presented here but has been expanded to include diverse DMSO oxidoreductases for direct comparison with the NarG/NxrA enzymes. Notice that both OP1 (green, blue, grey) and Gamma-type (red, orange) variants cluster within the cytoplasmically oriented Nar and Nxr enzymes. Six-hundred and ninety-seven NarG/NxrA proteins were identified from UniRef[63], and from those, 321 full-length sequences were selected to represent all the diverse clades. An additional 71 non-NarG/NxrA proteins, representative of the diverse enzymes of the DMSO superfamily were also included in the collection. The full-length amino acid sequences were aligned with Clustal Omega[61] and the phylogenetic tree was constructed by maximum likelihood and 1,000 bootstraps using RAxML[60]. The alignment length was 1,803 columns, out of which 31.2% were gaps or undetermined. Partial NarG sequences identified in the SAGs were placed on the tree using the EPA algorithm from RAxML[66]. The same collection of proteins was used to train the Rocker models and quantify the *narG* metagenomic fragments, and can be found in the enve-omics website (http://enve-omics.ce.gatech.edu/rocker/models). **b**, Alignment of NarG sequences from OMZ SAR11 with representative sequences from the DMSO superfamily of oxidoreductases. The protein motifs in the second and third panels are present in all functional Nar enzymes (NarG) and Nxr enzymes (NxrA) but not in closely related enzymes of the DMSO superfamily. The first panel shows the presence/absence of the TAT signal peptide (SRRSFLK), whose presence typically denotes a protein excreted to the outer membrane[40,41]. SAR11 NarG is instead oriented towards the cytoplasm (lack of TAT). The second panel shows the cysteine-rich motif typically found in the N terminus of the type-II DMSO superfamily oxidoreductases[75] and believed to enable the formation of a [4Fe–4S] cluster in these proteins[76]. The Asn in position 158 of the alignment is typically found in catalytic subunits of nitrite reductases and DMSO oxidoreductases (DmsA) but not in other DMSO family enzymes. The third panel shows the Gln(Q) and Thr(T) in positions 398 and 399 within the putative substrate entry channel of the protein, which differentiate the Nar proteins from all other oxidoreductases of the DMSO family[40].



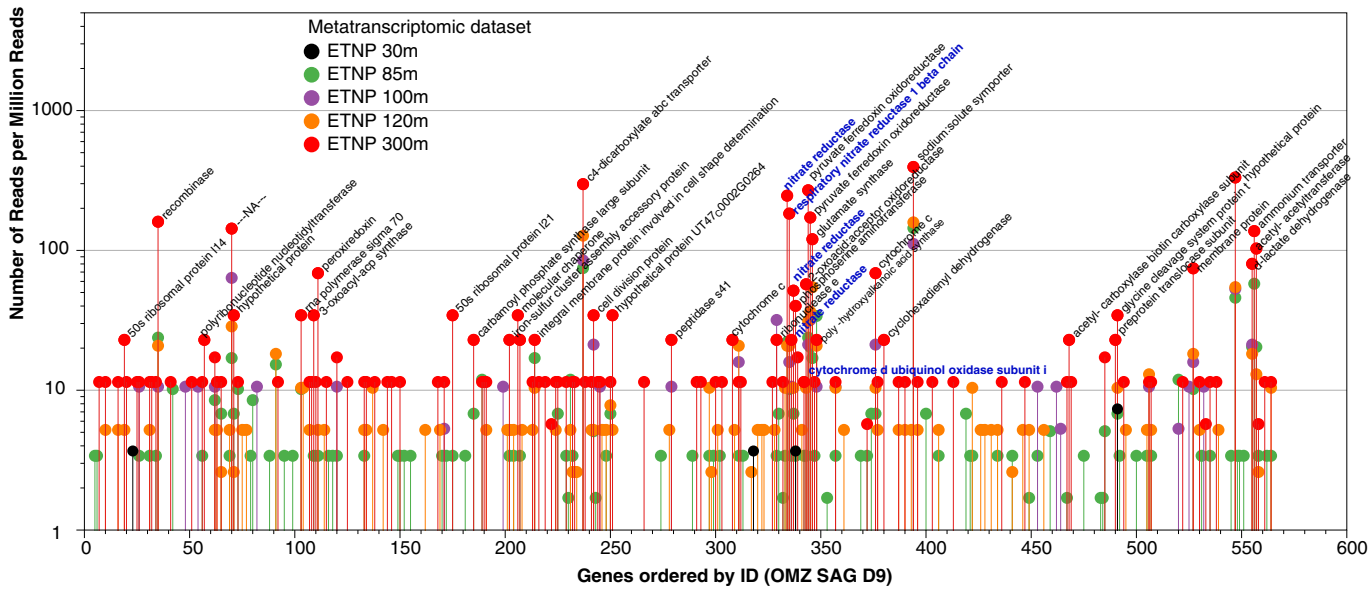
Extended Data Figure 5 | Functional characterization of the SAR11 *nar* operons in the *E. coli* heterologous expression system. a, Genotype of the *E. coli* triple mutant confirmed by whole-genome sequencing. The triple mutant lacks complete functional operons of all three NO_3^- reductase enzymes, and thus is incapable of NO_3^- reduction. **b**, Anaerobic growth of triple-mutant clones, complemented with the SAR11 *nar* operons. For each strain three independent clones were monitored, and data from the replicate growth curves were fitted into a logistic model. Shaded areas represent the 95% confidence intervals of optical density readings ($OD_{600\ nm}$) in the fitted logistic growth models. NO_3^- and NO_2^- were measured in parallel with ion chromatography. Note that the Gamma-type SAR11 operon complements the triple-mutant phenotype, growing anaerobically by reducing NO_3^- to NO_2^-. *E. coli* encodes functional nitrite reductases, thus the accumulated NO_2^- can be further reduced to ammonia, accounting for the non-stoichiometric NO_2^- production. **c**, Whole-cell NO_2^- production assays under aerobic conditions. Eight independent clones (columns A–H) of each type (C1–C5) were inoculated in Luria–Bertani (LB) broth supplemented with 30 mM NO_3^- and different isopropyl-β-D-thiogalactoside (IPTG) concentrations, and the well plate was incubated for 2 days at room temperature. Griess reagent was added, and development of pink colour indicated NO_2^- production.



Extended Data Figure 6 | Relative abundance of *narG* variants in ETNP OMZ metagenomes and metatranscriptomes and various other ocean metagenomes. a, Relative abundance and diversity of NarG/NxrA enzymes as revealed by phylogenetic placement of identified *narG* metagenomic reads (coloured pies). All identified short metagenomic *narG* reads from various oceanic metagenomes were placed within a reconstructed reference NarG tree to estimate the abundance of the different *narG* variants. The results of the placement are presented in five separate trees, based on the origin of the analysed metagenomic reads (ETSP metagenomes, ETNP metagenomes and metatranscriptomes, oxic bathypelagic and oxic surface metagenomes) for clarity. In each of the five trees, the coloured pies represent the abundance (normalized for data set size) of the short metagenomic reads clustering in the respective node. Specifically, the pie radius reflects read abundance as a percentage of the total *narG* genome equivalents identified (that is, number of *narG* reads compared to number of *rpoB* reads, normalized for gene length and total number of reads in each metagenome), with the size of grey pies representing the highest and lowest relative abundance, respectively. The reference tree is the same as in Fig. 3a. Scale bars represent substitutions per amino acid. Notice that the two *narG* variants affiliated with the SAR11 SAGs (highlighted in orange for the OP1 type and blue for the Gamma type) are only abundant in the metagenomes and metatranscriptomes from the OMZ, where they comprise more than 70% of the total *narG* read pool, as can also be observed in Fig. 3b and c. The number of *narG* reads of the OP1 or Gamma type are also given in Supplementary Table 1. **b**, qPCR-based abundance of SAR11-affiliated *narG* genes in the ETNP OMZ relative to NO_2^-, NO_3^- and O_2 concentrations and qPCR-based counts of 16S rRNA. Counts of total bacterial 16S rRNA, OP1-type *narG*, and Gamma-type *narG* genes at three stations (map on legend) west of Manzanillo, Mexico in May 2014. Map was created with Ocean Data View (http://odv.awi.de). All assays were performed in triplicates, and the bars represent s.e.m. Note that counts of OP1- and Gamma-type *narG* variants are probably underestimates given the observed microdiversity in the community (Extended Data Figs 2 and 7), and therefore there is a possibility that our primers did not match all OP1- and Gamma-type variants.



Extended Data Figure 7 | Diversity of OP1 and Gamma-type *narG* amino acid sequences in the ETNP OMZ metagenome. a, Phylogenies showing all full-length *narG* sequences recovered in the ETNP OMZ metagenomes (85, 100, 125, 300 m), as well as those from the SAR11 SAGs and corresponding *narG* reference sequences, with the left tree showing OP1-type variants and the right tree showing Gamma-type variants. NarG sequences are colour-coded based on the taxonomic classification of adjacent genes in the same metagenomic scaffolds, as show in Supplementary Table 6. **b**, Recruitment of metagenomic reads (predicted open reading frames) from the OMZ 300 m sample, against OP1- (left) or Gamma- (right) type *narG* sequences from the SAR11 SAGs. The metagenomic reads used for recruitment were identified as '*narG*' using the ROCker pipeline, and their identity was further confirmed by phylogenetic placement within the *narG* clade on a reference DMSO superfamily protein tree, to minimize non-specific recruitments in conserved protein regions. Note that based on this analysis, the OP1-type *narG* variants are highly diverse in the OMZ metagenome.



Extended Data Figure 8 | Transcriptional profile of predicted genes from the SAR11 OMZ SAG-D9. Transcriptomic reads with >99% identity matches were counted for each gene, and the counts were normalized for the data set size. Note that the *nar* operon genes are among the most actively transcribed in the ETNP 300 m OMZ sample.

ARTICLE

doi:10.1038/nature18943

An evolutionarily conserved pathway controls proteasome homeostasis

Adrien Rousseau[1] & Anne Bertolotti[1]

The proteasome is essential for the selective degradation of most cellular proteins, but how cells maintain adequate amounts of proteasome is unclear. Here we show that there is an evolutionarily conserved signalling pathway controlling proteasome homeostasis. Central to this pathway is TORC1, the inhibition of which induced all known yeast 19S regulatory particle assembly-chaperones (RACs), as well as proteasome subunits. Downstream of TORC1 inhibition, the yeast mitogen-activated protein kinase, Mpk1, acts to increase the supply of RACs and proteasome subunits under challenging conditions in order to maintain proteasomal degradation and cell viability. This adaptive pathway was evolutionarily conserved, with mTOR and ERK5 controlling the levels of the four mammalian RACs and proteasome abundance. Thus, the central growth and stress controllers, TORC1 and Mpk1/ERK5, endow cells with a rapid and vital adaptive response to adjust proteasome abundance in response to the rising needs of cells. Enhancing this pathway may be a useful therapeutic approach for diseases resulting from impaired proteasomal degradation.

Cell survival depends on adaptive signalling pathways to ensure that the supply of vital components matches fluctuating needs. The proteasome is essential for the selective degradation of most cellular proteins and thereby has a key role in most cellular processes[1–3]. Proteasome abundance is crucial for cell fitness, but how cells maintain adequate amounts of proteasome is unclear. Failure to degrade mutant or misfolded proteins causes diverse diseases, including devastating neurodegenerative diseases, which might potentially be prevented by increasing proteasome degradation[3]. Although the idea is attractive, increasing proteasome capacity remains a challenge. Thus, a better understanding of the mechanisms regulating proteasome abundance is required.

The proteasome is composed of 33 subunits assembled in two sub-complexes, the 20S core particle (CP), flanked at one or both ends by the 19S regulatory particle (RP) to form the 26S proteasome[2]. Proteasome assembly requires the assistance of proteasome assembly chaperones[4]. Four evolutionarily conserved 19S RACs: Nas2, Nas6, Hsm3 and Rpn14 in yeast, and p27 (also known as PSMD9), p28 (also known as PSMD10), S5b (also known as PSMD5) and Rpn14 (also known as PAAF1) in mammals are needed for regulatory particle assembly[5–9]. In addition, yeast cells have Adc17, a stress-inducible RAC, which is vital for cells to survive conditions, such as accumulation of misfolded proteins, which overwhelm the proteasome[10]. This suggests that cells have evolved adaptive signalling pathways to adjust proteasome assembly to arising needs, but how this is achieved is unknown.

TORC1 inhibition increases Adc17 and the proteasome

To determine how yeast cells maintain proteasome homeostasis, we decided to investigate the pathway regulating Adc17. Adc17 is upregulated by diverse stresses that impose a high burden on the proteasome, indicating that it is a component of an unknown generic stress response. Because Adc17 is induced by tunicamycin, an inducer of the unfolded protein response (UPR)[11], we deleted the UPR genes *IRE1* or *HAC1* (ref. 11). This prevented tunicamycin-mediated induction of the UPR marker Kar2, as expected[11], but not that of Adc17 (Fig. 1a), indicating that *ADC17* was not a UPR target gene. We tested Adc17 induction by tunicamycin in mutants thought to regulate Adc17 from a genome-wide regulation study[12], and found that deletion of *SFP1* abolished Adc17 but not Kar2 induction by tunicamycin (Fig. 1b). Adc17 induction by tunicamycin was higher in a strain carrying a hypomorphic allele of *MRS6*, a negative regulator of Sfp1 (Extended Data Fig. 1a). Sfp1 is a stress- and nutrient-sensitive regulator of cell growth with dual function[13–15]. Under optimal growth conditions, Sfp1 is located in the nucleus and can activate transcription of ribosomal protein genes, but it re-localizes to the cytosol upon stress[13,14]. Sfp1 is activated by TORC1, and in turn negatively regulates TORC1 signalling, as a feedback mechanism[15]. In the absence of Spf1, TORC1 is hyperactive[15]. Thus, *SFP1* deletion could prevent Adc17 induction directly or by over-activating TORC1. Adc17 induction by tunicamycin (Fig. 1a, b) coincided with Sfp1 re-localization from the nucleus to the cytosol (Extended Data Fig. 1b), suggesting that Sfp1 may regulate Adc17 not directly, but instead indirectly through TORC1. Tunicamycin inhibits TORC1 signalling[15], as observed (Fig. 1c) with the phosphorylation of the TORC1 effector Sch9 (ref. 16). In the absence of Sfp1, TORC1 was hyperactive[15] (Fig. 1c), and it remained active during tunicamycin-mediated stress, while Adc17 induction was abolished (Fig. 1c), suggesting that Sfp1 regulated Adc17 via TORC1. Confirming this, rapamycin, a selective inhibitor of TORC1 (ref. 17) induced Adc17 (Fig. 1d). Deletion of *SFP1* abolished induction of Adc17 by tunicamycin but not by rapamycin (Fig. 1e) because *SFP1* deletion affected Adc17 expression by hyperactivating TORC1 (Fig. 1f). To confirm this using a genetic approach, we examined Adc17 regulation in the thermosensitive *kog1-1* mutant. Kog1 (Fig. 1f) is the yeast homologue of Raptor, a subunit of TORC1 (ref. 18). Inactivation of *KOG1* inhibited TORC1, as expected[18], and induced Adc17 (Fig. 1g), indicating that selective TORC1 inhibition induces Adc17. We investigated whether rapamycin increased proteasome abundance. Consistent with our previous results for tunicamycin[10], proteasome levels increased by more than twofold after 3 h of rapamycin treatment (Fig. 1h, i). Thus, inhibition of the central stress and growth controller, TORC1, increases abundance of Adc17 and of the proteasome in yeast.

The MAPK Mpk1 induces Adc17

TORC1 integrates multiple signalling pathways[17,19]. We searched for the pathway downstream of TORC1 controlling Adc17 and proteasome abundance. Adc17 is not a UPR gene (Fig. 1a), but *adc17*Δ cells are sensitive to tunicamycin-mediated stress[10]. Therefore, we examined

[1]MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge, CB2 0QH, UK.

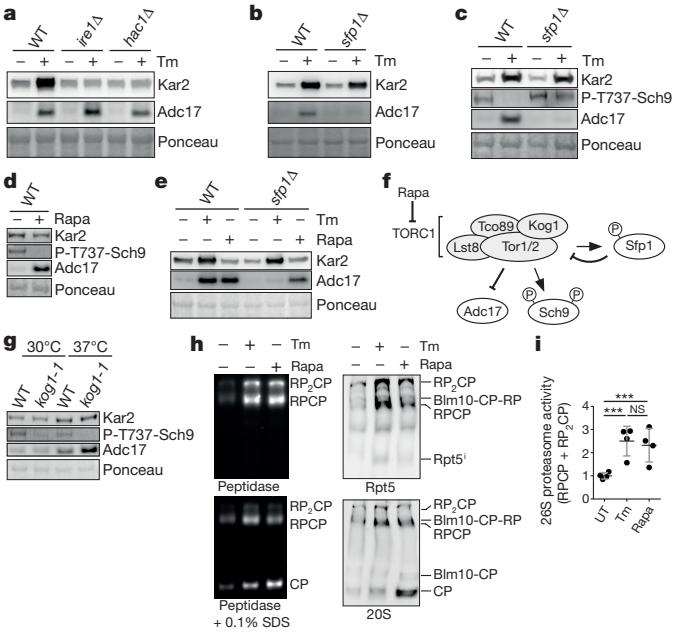


Figure 1 | TORC1 inhibition induces the proteasome assembly chaperone Adc17 and increases proteasome levels. a–c, Immunoblots of lysates from yeast cells treated ± tunicamycin (Tm) for 4 h. Here and elsewhere, unless specified otherwise, the tunicamycin treatment was ± 5 μg ml^{-1}. WT, wild type. **d**, Immunoblots of lysates from yeast cells treated ± rapamycin (Rapa) for 4 h. Here and elsewhere, unless specified otherwise, the rapamycin treatment was ± 0.2 μg ml^{-1}. **e**, Immunoblots of yeast cell lysates after treatment ± tunicamycin or rapamycin for 4 h. **f**, Schematic depicting the relationship between Sfp1, TORC1 and Adc17. **g**, Immunoblots from yeast cells cultured at 30 °C or 37 °C for 4 h. **h**, Native polyacrylamide gel electrophoresis (PAGE) (4.2%) of yeast extracts from cells treated with tunicamycin or rapamycin, monitored by the fluorogenic substrate Suc-LLVY–AMC and by immunoblots. **i**, Quantification of the 26S proteasome activity (RPCP and RP_2CP) in four independent experiments such as the one shown in **h**. Data are mean ± s.d.; $n = 4$ biological replicates. ***$P \leq 0.001$; NS, not significant (one-way ANOVA).

non-UPR mutants sensitive to tunicamycin. The mitogen-activated protein kinases (MAPKs) Hog1 and Mpk1 were important for tunicamycin-stress survival in yeast (Fig. 2a), as expected[20], unlike the other MAPKs Fus3, Kss1 and Smk1 (Fig. 2a); Hog1 being advantageous and Mpk1 essential for stress survival (Fig. 2b). Adc17 induction by tunicamycin was compromised in *HOG1* deleted cells and abolished in cells lacking a functional allele of *MPK1* (Fig. 2c and Extended Data Fig. 2a, b) revealing a perfect correlation between tunicamycin stress-resistance and Adc17 induction. Genetic interaction studies showed that overexpression of *HOG1* failed to restore tunicamycin resistance and Adc17 induction in *mpk1*Δ cells (Fig. 2d, e), while overexpression of *MPK1* increased both tunicamycin resistance and Adc17 induction in *hog1*Δ cells (Fig. 2f, g). Thus, signalling through Mpk1 is required for Adc17 induction and tunicamycin survival.

We examined if Mpk1 was required for Adc17 induction by rapamycin. *MPK1* is negatively regulated by TORC1 and essential for rapamycin survival[21,22]. Unlike the other MAPK, Mpk1 was essential for both cell viability and Adc17 induction in the presence of rapamycin (Extended Data Fig. 2c–e). *HOG1* contributed to Adc17 upregulation by tunicamycin (Fig. 2c) but not by rapamycin (Extended Data Fig. 2d). *HOG1* was dispensable for survival in the presence of rapamycin (Extended Data Fig. 2c). Thus, induction of Adc17 and rapamycin-resistance are perfectly correlated (Extended Data Fig. 2c, d). In agreement with a previous study[23], the levels of Mpk1 increased in response to tunicamycin treatment (Extended Data Fig. 2d), but this increase was markedly attenuated in *hog1*Δ cells (Extended Data Fig. 2d). Thus, one key function of Hog1 is to regulate Mpk1 levels (Fig. 2h), providing an explanation for why Mpk1 overexpression in *hog1*Δ cells rescued tunicamycin-resistance and Adc17 induction (Fig. 2f, g). Over time, both Mpk1 phosphorylation and abundance were increased by tunicamycin and rapamycin treatment and this preceded Adc17 induction (Extended Data Fig. 3a, b). Bck1, Mkk1 and Mkk2, three kinases that are upstream of Mpk1 (ref. 24), were also required for Adc17 induction by tunicamycin and rapamycin treatment (Extended Data Fig. 3c, d). Congo red, a cell-wall-damaging agent and known inducer of the Mpk1 MAPK pathway[24] also induced Adc17, in a Mpk1-dependent manner (Extended Data Fig. 3e). These results indicate that diverse challenges inhibiting TORC1 signal to the Mpk1 MAPK to induce the proteasome assembly chaperone Adc17.

Mpk1 is a master regulator of the proteasome

We focused on Mpk1 because it is essential for Adc17 induction (Fig. 3a) and examined whether Mpk1 regulated proteasome abundance. Deleting *MPK1* completely abolished the tunicamycin- or rapamycin-induced increase of 26S proteasomes while increasing the abundance of the free core particles (Fig. 3b–d). This defect is symptomatic of regulatory particle assembly defects[6–8], and a hallmark of *adc17*Δ cells in response to stress[10]. However, the *mpk1*Δ cells (Fig. 3b–d) appeared more severely affected than *adc17*Δ cells[10], suggesting that other MPK1-regulated factors assist regulatory particle assembly. We found that all the known yeast RACs: Nas2, Nas6, Hsm3 and Rpn14 were induced by treatment with tunicamycin, rapamycin or Congo red in wild-type cells (Fig. 3e and Extended Data Fig. 3f). Genetic inactivation of TORC1 in *kog1-1* cells also induced all RACs at the non-permissive temperature (Fig. 3f). Induction of all yeast RACs by tunicamycin and rapamycin was abolished in *mpk1*Δ, *bck1*Δ and *mkk1/2*Δ cells (Fig. 3g and Extended Data Fig. 3g, h). Overexpression of different combinations of three RACs markedly improved tunicamycin resistance in *mpk1*Δ cells (Extended Data Fig. 4a). Conversely, the deletion of three RACs severely impaired cell viability in the presence of rapamycin (Extended Data Fig. 4b). Thus, regulating the expression of RACs is a key function of Mpk1. These results reveal that downstream of TORC1 inhibition, signalling through the Mpk1 MAPK pathway coordinates the induction of all RACs to control proteasome abundance and viability upon various stresses.

Tunicamycin and rapamycin increased 26S abundance in wild-type cells and increased free core particles in *mpk1*Δ cells (Fig. 3b), suggesting that core particle assembly might also be regulated. We analysed the levels of the core particle assembly chaperones proteasome biogenesis-associated (Pba)1–4 (refs 25, 26) after tunicamycin treatment, the most potent inducer of core particles in *mpk1*Δ cells (Fig. 3b, d). In wild-type cells, tunicamycin treatment increased the level of Pba1 and Pba2 but not the level of Pba3 and Pba4 (Extended Data Fig. 5a–d). Thus, the increase in core particles was accompanied by an increase of the assembly chaperones Pba1 and Pba2. This increase was unaltered upon *MPK1* deletion (Extended Data Fig. 5a–d). This demonstrates that Pba1 and Pba2 are upregulated by the stress caused by tunicamycin treatment and their regulation is independent of Mpk1. The mechanism of Mpk1-independent regulation of Pba1 and Pba2 will be an important topic for future study.

We examined the regulation of proteasome subunits. Both tunicamycin and rapamycin treatment increased the levels of proteasome subunits, and this increase required Rpn4, the transcription factor controlling expression of proteasome subunits[27] (Extended Data Fig. 6a, b). Rpn4 increased upon tunicamycin or rapamycin treatment (Extended Data Fig. 6c). In contrast, Adc17 is upregulated independently of Rpn4 upon diverse stresses (ref. 10), and all yeast RACs show this same pattern of regulation (Extended Data Fig. 6b). Upregulation of proteasome subunits depends on Rpn4, and upregulation of all known RACs is independent of Rpn4. Deletion of *MPK1* completely abrogated the tunicamycin- and rapamycin-induced upregulation of proteasome subunits, indicating that Mpk1 is a master regulator of proteasome homeostasis (Fig. 4a and Extended Data Fig. 6d).

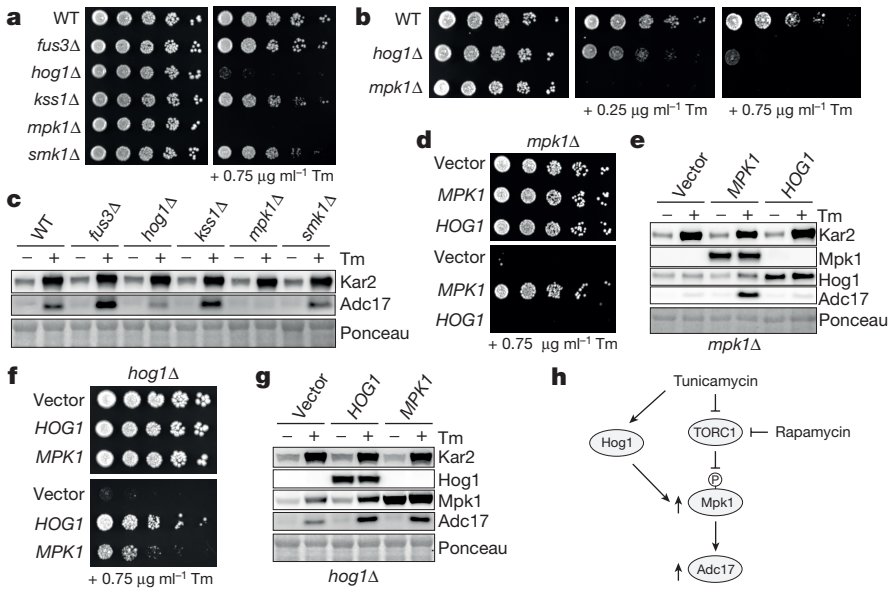


Figure 2 | The MAPK Mpk1 is a master regulator of the stress-inducible proteasome assembly chaperone Adc17. a, **b**, Cells spotted in a sixfold dilution and grown for 3 days on plates ± tunicamycin. **c**, Immunoblots from yeast cells grown ± tunicamycin for 4 h. **d**, **f**, Cells transformed with empty vector or with *MPK1* or *HOG1* spotted in a sixfold dilution and grown on plates ± 0.75 μg ml^{-1} tunicamycin for 3 days. **e**, **g**, Immunoblots of lysates from yeast cells grown ± tunicamycin for 4 h. **h**, Schematic of the Adc17 signalling pathway.

We identified a weak genetic interaction between *RPN4* and *MPK1*, and found that both were required for survival in response to tunicamycin treatment (Extended Data Fig. 6e, f). Tunicamycin and rapamycin increased Rpn4 levels to wild-type levels in *mpk1Δ* cells (Extended Data Fig. 6g), suggesting that Mpk1 is acting downstream of the transcription factor Rpn4, possibly post-transcriptionally. At the protein level, *MPK1* deletion completely abrogated the induction of proteasome subunits and RACs by rapamycin treatment (Fig. 4a). At the mRNA level, rapamycin only modestly, yet reproducibly, increased abundance of RACs and proteasome subunits mRNA (Fig. 4b), and this increase was similar in wild-type and *mpk1Δ* cells (Fig. 4b). Rpn4 induction was similar in both strains (Extended Data Fig. 6g). Blocking the synthesis of new proteins with cycloheximide for 4 h did not change the abundance of proteasome subunits and RACs, indicating that they

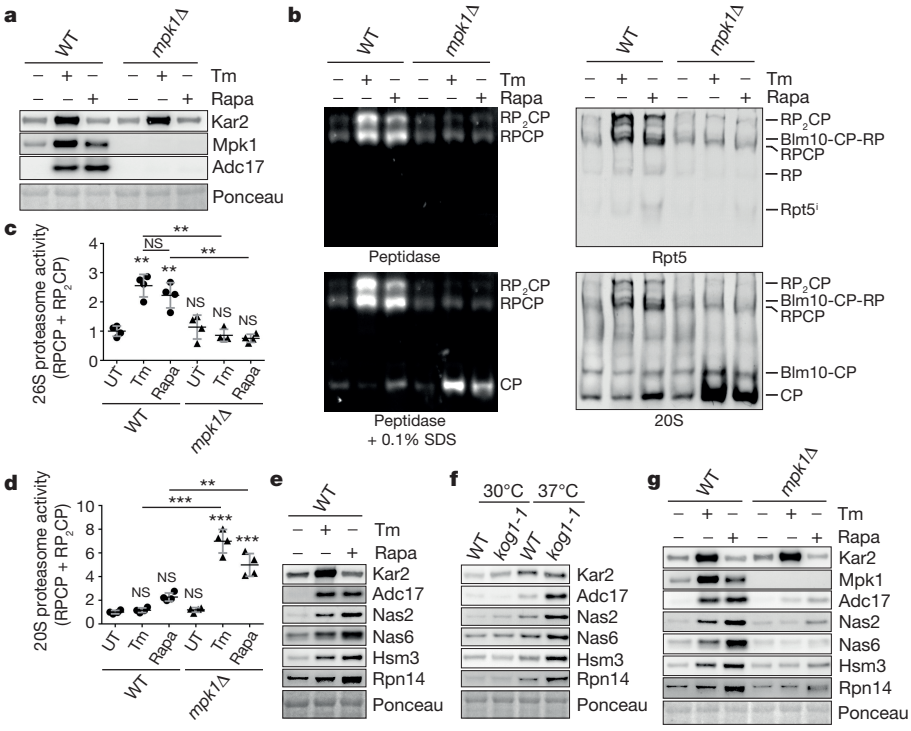


Figure 3 | Mpk1 coordinates the expression of all yeast RACs to control proteasome abundance. a, Immunoblots of lysates from yeast cells cultured ± tunicamycin or rapamycin for 4 h. **b**, Native PAGE (4.2%) of yeast cells cultured ± tunicamycin or rapamycin, monitored by Suc-LLVY-AMC and by immunoblots. Rpt5^{i} (Rpt5 intermediates). **c**, **d**, Quantifications from experiments as in **b**. Data are mean ± s.d. of four biological replicates. $^{**}P \leq 0.01$; $^{***}P \leq 0.001$; NS, not significant (two-way ANOVA). **e**, Immunoblots from lysates of yeast cells cultured ± tunicamycin or rapamycin for 4 h. **f**, Immunoblots from lysates of yeast cells cultured at 30 °C or 37 °C for 4 h. **g**, Immunoblots from lysates of yeast cells cultured ± tunicamycin or rapamycin for 4 h.

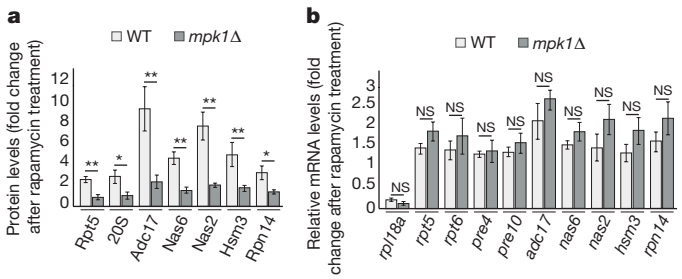


Figure 4 | Post-transcriptional control of RAC and proteasome subunit abundance by Mpk1. a, Relative abundance of the indicated proteins in yeast cells treated with rapamycin for 4 h relative to untreated cells. **b**, Relative abundance of the indicated mRNA in yeast cells treated with rapamycin for 2 h relative to untreated cells. *rpl18a* was used as a control. **a**, **b**, Data are mean ± s.d.; $n = 3$ biological replicates. $*P \leq 0.05$, $**P \leq 0.01$; NS, not significant (two-way ANOVA).

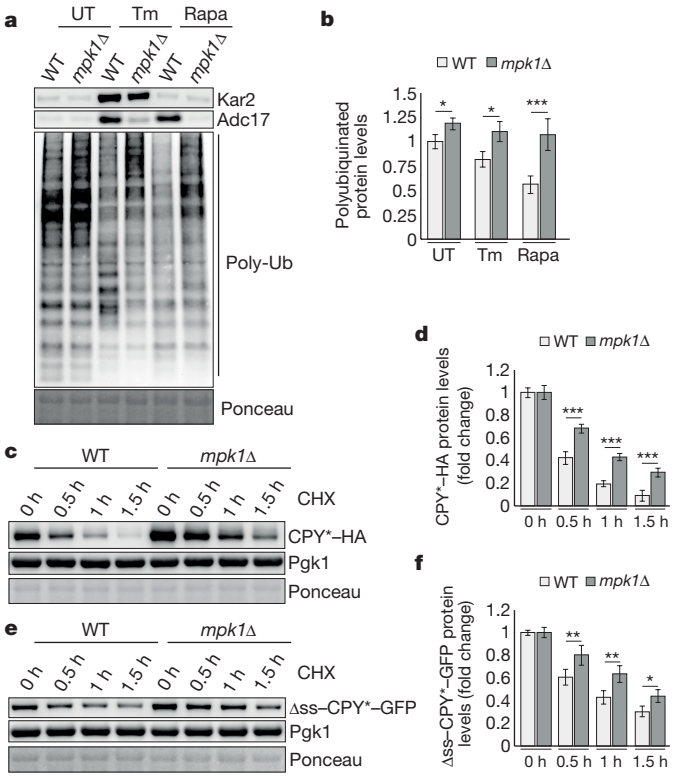


Figure 5 | Mpk1 adjusts proteasome degradation to match the needs of the cell. a, Immunoblots of lysates of yeast cells cultured ± tunicamycin or rapamycin for 4 h. Poly-Ub: polyubiquitinated conjugates. **c**, **e**, Immunoblots from lysates of yeast cells expressing CPY*–HA (**c**) or Δss-CPY*–GFP (**e**) treated with 35 μg ml^{-1} cycloheximide (CHX) for the indicated time. **b**, **d** and **f**, show quantification of **a**, **c** and **e**, respectively. Data are mean ± s.d.; $n = 4$ (**b**) and $n = 3$ (**d**, **f**) biological replicates. $*P \leq 0.05$; $**P \leq 0.01$; $***P \leq 0.001$; NS, not significant (two-way ANOVA).

were stable over this time period (Extended Data Fig. 6h, lanes 1 and 4). Likewise, the stability of proteasome subunits and RACs appeared similar in *mpk1Δ* cells and wild-type cells (Extended Data Fig. 6i). However, cycloheximide completely blocked induction of proteasome subunits and RACs by tunicamycin and rapamycin in wild-type cells (Extended Data Fig. 6h). Together these results reveal that the MAPK Mpk1 coordinates the translation of proteasome subunits and RACs to provide the increased proteasome abundance required to sustain cell viability.

Mpk1 adapts proteasome degradation to rising needs

We analysed the consequences of the MPK1-dependent increase of proteasome abundance on protein degradation. Polyubiquitinated conjugates represent a hallmark of impaired proteasomal degradation and were slightly elevated in *mpk1Δ* cells compared to wild type (Fig. 5a, b). This defect was exacerbated upon tunicamycin or rapamycin treatment (Fig. 5a, b), suggesting impaired proteasomal degradation, and providing an explanation for why *mpk1Δ* cells failed to survive tunicamycin (Fig. 2a) or rapamycin treatment (Extended Data Fig. 2c).

We examined the degradation of diverse proteasome reporter substrates. The metastable Ura3-3 reporter[28] was rapidly degraded in wild-type cells cultured at 37 °C, but not in cells harbouring a thermosensitive mutation in the proteasome subunit Rpt4 (Extended Data Fig. 7a, b). Similarly, the degradation of the reporter substrate was strikingly compromised in *mpk1Δ* cells (Extended Data Fig. 7c, d). The degradation of the two well-characterized proteasome reporter substrates, CPY*–HA and Δss-CPY*–GFP, which are localized in the endoplasmic reticulum and in the cytosol, respectively[29,30], was also compromised in *mpk1Δ* cells (Fig. 5c–f). The protein degradation defect of *mpk1Δ* cells was more pronounced in cells challenged with tunicamycin and rapamycin treatment (Extended Data Fig. 7e–l). Together with the previous findings, this demonstrates that Mpk1 maintains adequate levels of proteasome required to sustain protein degradation and cell viability under challenging conditions.

Evolutionary conservation of proteasome regulation

Four RACs are evolutionarily conserved with p27, p28, S5b and Rpn14 being human orthologues of the yeast Nas2, Nas6, Hsm3 and Rpn14, respectively[5–9]. We investigated whether the TORC1 and Mpk1 regulation of RACs was evolutionarily conserved. Inhibition of mTOR by Torin-1 rapidly increased the levels of all mammalian RACs (Fig. 6a, b), similar to what was found in the experiments in yeast (Fig. 3e, f). mTOR inhibition resulting from nutrient starvation also increased the RACs (Extended Data Fig. 8a, b). As in yeast, the concerted increase of the RACs was accompanied by an upregulation of proteasome subunits (Fig. 6a, b), and resulted in an increase in the levels of 26S proteasome (Fig. 6c, d and Extended Data Fig. 8c, d). This response was acute, with a rapid return to basal levels (Fig. 6a–d). As previously reported[31], singly capped proteasome RPCP (CP, core particle; RP, regulatory particle) was more abundant than doubly capped proteasome RP_2CP in mammalian cells (Fig. 6c).

Conversely, medium replenishment to increase nutrient supply and activate mTORC1 had the opposite effect resulting in S6K1 phosphorylation (Extended Data Fig. 9a) and decreasing abundance of both RACs (Extended Data Fig. 9a, b) and proteasome (Extended Data Fig. 9c, d). Rapamycin, a selective mTORC1 inhibitor, also acutely and transiently induced the RACs as well as proteasome subunits (Extended Data Fig. 10), confirming that, as in yeast, mTORC1 controls proteasome homeostasis. We wondered whether ERK5 (also known as MAPK7) (ref. 32), the mammalian orthologue of Mpk1, also regulates proteasome abundance. ERK5 overexpression in yeast rescued tunicamycin resistance in *mpk1Δ* cells (Fig. 6e). Knocking down ERK5 with short interfering RNA (siRNA) in human cells resulted in a decrease of the four mammalian RACs p27, p28, S5b and Rpn14 (Fig. 6f, g), as well as the 26S proteasome (Fig. 6h, i). Thus, mammalian ERK5, like yeast Mpk1, controls RACs and thereby acts as a switch to control proteasome abundance.

Discussion

Here we report a general and evolutionarily conserved homeostatic response that increases proteasome abundance as needed, through the coordinated upregulation of regulatory particle assembly chaperones and proteasome subunits. The master regulators of growth and stress, TORC1 and Mpk1/ERK5 are central to this response. Consistent with the general principle of homeostatic responses, we observed that proteasome increase is an acute and rapidly reversible response. Trying to identify the other components of this proteasome homeostatic

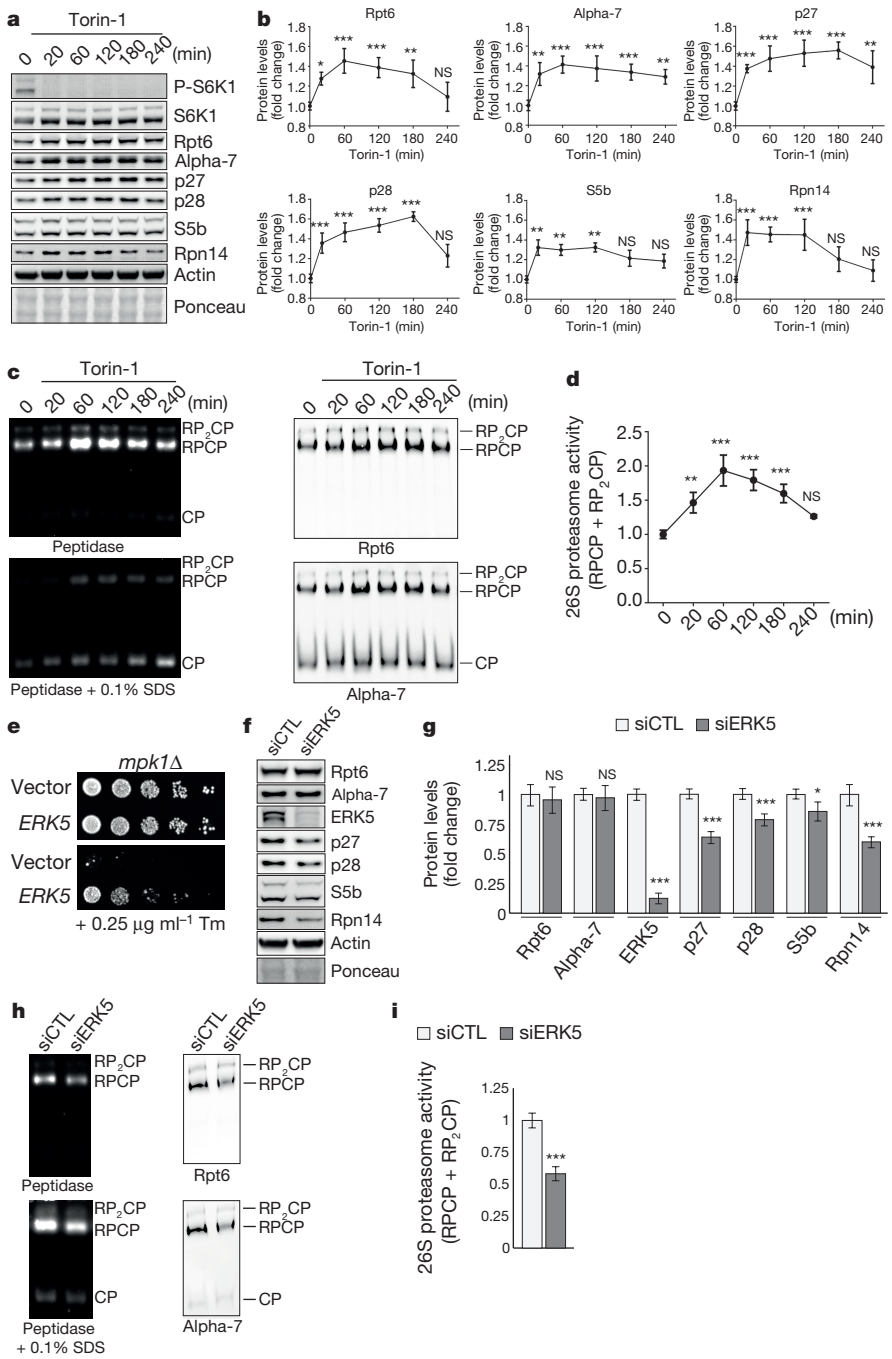


Figure 6 | Evolutionary conservation of the pathway controlling RACs and proteasome abundance. a, **b**, Immunoblots (**a**) and quantifications (**b**) of the indicated proteins in lysates of HeLa cells treated with 250 nM Torin-1 for the indicated time. **c**, **d**, Native PAGE (4.2%) (**c**) and quantifications (**d**) of HeLa cell lysates following treatment as in **a** and revealed with Suc-LLVY–AMC and by immunoblots. **e**, *mpk1Δ* cells transformed with a plasmid encoding the human ERK5 or an empty vector were spotted in a sixfold dilution and grown on plates ± tunicamycin for 3 days. **f**, **g**, Immunoblots (**f**) and quantifications (**g**) of the indicated proteins in lysates of HeLa cells 3 days after transfection with a non-target siRNA (siCTL) or a siRNA targeting ERK5 (siERK5). **h**, **i**, Native PAGE (4.2%) (**h**) and quantifications (**i**) of HeLa cell extracts 3 days after transfection with siCTL or siERK5 monitored by Suc-LLVY–AMC or by immunoblots. **b**, **d**, **g**, **i**, Data are mean ± s.d; $n = 3$ biological replicates. $^{*}P \leq 0.05$; $^{**}P \leq 0.01$; $^{***}P \leq 0.001$; NS, not significant (**b**, **d**, one-way ANOVA; **g**, two-way ANOVA; **i**, two-tailed Student's *t*-test).

response, in particular the mechanisms regulating 20S assembly and determining how proteasome levels return to baseline after an acute increase will be the subject of future studies.

Our results also provide a framework for resolving inconsistences in previous observations. It was reported that when cultured in absence of serum, proteasomal degradation is increased in cells lacking Tsc2, a negative regulator of TORC1 (ref. 33). In contrast to this, a recent study reported that mTOR inhibition activates proteasomal degradation by a mechanism proposed to be driven by increased ubiquitination[34]. Considering this in light of our results, it may be the adaptive response to the stress resulting from the lack of Tsc2 combined with serum starvation that increases proteasomal degradation in *Tsc2*$^{-/-}$ cells, rather than *Tsc2* deletion per se.

In line with our findings is the well-established notion that mTOR activation enhances anabolic processes and represses catabolic processes[19,35]. mTORC1 is known to repress autophagy. We show here that TORC1 restricts proteasome abundance and this is rapidly alleviated upon TORC1 inhibition. Therefore, the same controller TORC1

restricts the abundance of the two cellular proteolytic systems, the proteasome and autophagy. Our findings integrate the regulation of proteasome assembly and abundance with growth and cellular metabolism, and suggest that the increased proteasome capacity resulting from TORC1 inhibition may also contribute to the benefit of the widely used TORC1 inhibitors.

The current prevailing view is that protein degradation is largely regulated at the level of ubiquitination. Here we demonstrate that modulating proteasome abundance is an important component of regulation of proteasomal degradation. Adapting proteasome abundance is vital to cope with overwhelming cellular needs, implying that proteasome abundance can be rate limiting under critical conditions. The evolutionary conservation of the TORC1 and Mpk1/ERK5 pathway controlling proteasome abundance further highlights the importance of this regulation.

The pathway identified here can be used as a unique switch to increase proteasome assembly and abundance on demand. Because many human diseases are associated with accumulation of misfolded proteins, increasing proteasome abundance by manipulating the switches identified here could be used as a generic strategy to reduce the burden of misfolded proteins that accumulate in such age-related diseases.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 January; accepted 20 June 2016.
Published online 27 July 2016.

1. Goldberg, A. L. Functions of the proteasome: from protein degradation and immune surveillance to cancer therapy. *Biochem. Soc. Trans.* **35,** 12–17 (2007).
2. Finley, D. Recognition and processing of ubiquitin–protein conjugates by the proteasome. *Annu. Rev. Biochem.* **78,** 477–513 (2009).
3. Tanaka, K., Mizushima, T. & Saeki, Y. The proteasome: molecular machinery and pathophysiological roles. *Biol. Chem.* **393,** 217–234 (2012).
4. Tomko, R. J. Jr & Hochstrasser, M. Molecular architecture and assembly of the eukaryotic proteasome. *Annu. Rev. Biochem.* **82,** 415–445 (2013).
5. Le Tallec, B., Barrault, M. B., Guérois, R., Carré, T. & Peyroche, A. Hsm3/S5b participates in the assembly pathway of the 19S regulatory particle of the proteasome. *Mol. Cell* **33,** 389–399 (2009).
6. Saeki, Y., Toh-E, A., Kudo, T., Kawamura, H. & Tanaka, K. Multiple proteasome-interacting proteins assist the assembly of the yeast 19S regulatory particle. *Cell* **137,** 900–913 (2009).
7. Funakoshi, M., Tomko, R. J. Jr, Kobayashi, H. & Hochstrasser, M. Multiple assembly chaperones govern biogenesis of the proteasome regulatory particle base. *Cell* **137,** 887–899 (2009).
8. Roelofs, J. *et al.* Chaperone-mediated pathway of proteasome regulatory particle assembly. *Nature* **459,** 861–865 (2009).
9. Kaneko, T. *et al.* Assembly pathway of the mammalian proteasome base subcomplex is mediated by multiple specific chaperones. *Cell* **137,** 914–925 (2009).
10. Hanssum, A. *et al.* An inducible chaperone adapts proteasome assembly to stress. *Mol. Cell* **55,** 566–577 (2014).
11. Wiseman, R. L., Haynes, C. M. & Ron, D. SnapShot: The unfolded protein response. *Cell* **140,** 590–590.e2 (2010).
12. Venters, B. J. *et al.* A comprehensive genomic binding map of gene and chromatin regulatory proteins in *Saccharomyces*. *Mol. Cell* **41,** 480–492 (2011).
13. Marion, R. M. *et al.* Sfp1 is a stress- and nutrient-sensitive regulator of ribosomal protein gene expression. *Proc. Natl Acad. Sci. USA* **101,** 14315–14322 (2004).
14. Jorgensen, P. *et al.* A dynamic transcriptional network communicates growth potential to ribosome synthesis and critical cell size. *Genes Dev.* **18,** 2491–2505 (2004).
15. Lempiäinen, H. *et al.* Sfp1 interaction with TORC1 and Mrs6 reveals feedback regulation on TOR signaling. *Mol. Cell* **33,** 704–716 (2009).
16. Takahara, T. & Maeda, T. Transient sequestration of TORC1 into stress granules during heat stress. *Mol. Cell* **47,** 242–252 (2012).
17. Soulard, A. & Hall, M. N. SnapShot: mTOR signaling. *Cell* **129,** 434.e1–434.e2 (2007).
18. Loewith, R. & Hall, M. N. Target of rapamycin (TOR) in nutrient signaling and growth control. *Genetics* **189,** 1177–1201 (2011).
19. Zoncu, R., Efeyan, A. & Sabatini, D. M. mTOR: from growth signal integration to cancer, diabetes and ageing. *Nature Rev. Mol. Cell Biol.* **12,** 21–35 (2011).
20. Bonilla, M. & Cunningham, K. W. Mitogen-activated protein kinase stimulation of Ca^{2+} signaling is required for survival of endoplasmic reticulum stress in yeast. *Mol. Biol. Cell* **14,** 4296–4305 (2003).
21. Krause, S. A. & Gray, J. V. The protein kinase C pathway is required for viability in quiescence in *Saccharomyces cerevisiae*. *Curr. Biol.* **12,** 588–593 (2002).
22. Torres, J., Di Como, C. J., Herrero, E. & De La Torre-Ruiz, M. A. Regulation of the cell integrity pathway by rapamycin-sensitive TOR function in budding yeast. *J. Biol. Chem.* **277,** 43495–43504 (2002).
23. Babour, A., Bicknell, A. A., Tourtellotte, J. & Niwa, M. A surveillance pathway monitors the fitness of the endoplasmic reticulum to control its inheritance. *Cell* **142,** 256–269 (2010).
24. Levin, D. E. Regulation of cell wall biogenesis in *Saccharomyces cerevisiae*: the cell wall integrity signaling pathway. *Genetics* **189,** 1145–1175 (2011).
25. Hirano, Y. *et al.* A heterodimeric complex that promotes the assembly of mammalian 20S proteasomes. *Nature* **437,** 1381–1385 (2005).
26. Le Tallec, B. *et al.* 20S proteasome assembly is orchestrated by two distinct pairs of chaperones in yeast and in mammals. *Mol. Cell* **27,** 660–674 (2007).
27. Xie, Y. & Varshavsky, A. RPN4 is a ligand, substrate, and transcriptional regulator of the 26S proteasome: a negative feedback circuit. *Proc. Natl Acad. Sci. USA* **98,** 3056–3061 (2001).
28. Suraweera, A., Münch, C., Hanssum, A. & Bertolotti, A. Failure of amino acid homeostasis causes cell death following proteasome inhibition. *Mol. Cell* **48,** 242–253 (2012).
29. Hiller, M. M., Finger, A., Schweiger, M. & Wolf, D. H. ER degradation of a misfolded luminal protein by the cytosolic ubiquitin-proteasome pathway. *Science* **273,** 1725–1728 (1996).
30. Medicherla, B., Kostova, Z., Schaefer, A. & Wolf, D. H. A genomic screen identifies Dsk2p and Rad23p as essential components of ER-associated degradation. *EMBO Rep.* **5,** 692–697 (2004).
31. Asano, S. *et al.* A molecular census of 26S proteasomes in intact neurons. *Science* **347,** 439–442 (2015).
32. Truman, A. W. *et al.* Expressed in the yeast *Saccharomyces cerevisiae*, human ERK5 is a client of the Hsp90 chaperone that complements loss of the Slt2p (Mpk1p) cell integrity stress-activated protein kinase. *Eukaryot. Cell* **5,** 1914–1924 (2006).
33. Zhang, Y. *et al.* Coordinated regulation of protein synthesis and degradation by mTORC1. *Nature* **513,** 440–443 (2014).
34. Zhao, J., Zhai, B., Gygi, S. P. & Goldberg, A. L. mTOR inhibition activates overall protein degradation by the ubiquitin proteasome system as well as by autophagy. *Proc. Natl Acad. Sci. USA* **112,** 15790–15797 (2015).
35. Albert, V. & Hall, M. N. mTOR signaling in cellular and organismal energetics. *Curr. Opin. Cell Biol.* **33,** 55–66 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank Y. Lee and M. Hochstrasser for the kind gift of Nas2, Nas6, Hsm3 and Rpn14 antibodies; D. H. Wolf for CPY*–HA and Δss-CPY*–GFP constructs; T. Maeda for the P-Sch9 antibody; and members of the Bertolotti laboratory for discussion. A.B. is an honorary fellow of the University of Cambridge Clinical Neurosciences Department. This work was supported by the Medical Research Council (UK) MC_U105185860. A.R. is supported by an EMBO long-term fellowship.

Author Contributions A.R. designed, performed and analysed all experiments, prepared the figures and helped with the manuscript. A.B. designed and supervised the study and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.B. (aberto@mrc-lmb.cam.ac.uk).

Reviewer Information *Nature* thanks S. Murata and D. Sabatini and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Yeast strains, plasmids and growth assays. Gene-deletion mutants and their isogenic wild-type strain (BY4741) were grown in YPD medium according to standard protocols[36]. To assess growth phenotypes, exponentially growing liquid cultures expressing the indicated genes were equilibrated to an OD_{600} (A_{600nm}) of 0.2, and 4 µl samples were spotted in serial dilutions (1:6) onto YPD or selective media as required. Plates were incubated at 30 °C for 3 days. To assess tunicamycin and rapamycin sensitivity, cells were spotted on plates supplemented with tunicamycin (0.25 µg ml^{-1} or 0.75 µg ml^{-1}, as indicated) or rapamycin (20 ng ml^{-1}). Yeast strains and plasmids used in this study are presented in Supplementary Tables 1 and 2, respectively.

Tunicamycin (Sigma-Aldrich; 2.5 mg ml^{-1} stock) aliquots were stored at −20 °C and used within three months. Rapamycin (Sigma-Aldrich; 1 mM in DMSO) and Torin-1 (Santa Cruz Biotechnology; 1 mM in DMSO) aliquots were stored at −80 °C and used within a month. Cycloheximide (Sigma-Aldrich; 35 mg ml^{-1} in ethanol) was used at 35 µg ml^{-1} final concentration to inhibit translation in yeast.

Immunoblot analyses in yeast. 10 ml of exponentially growing cells adjusted to an OD_{600} of 0.2 were treated with 5 µg ml^{-1} tunicamycin (Tm), 0.2 µg ml^{-1}rapamycin, 50 µg ml^{-1}Congo red (CR) or DMSO for 4 h at 30 °C. Cells were harvested by centrifugation at 8,000*g* for 30 s at 4 °C, pre-treated with 2 M LiAc and then 0.4 M NaOH for 5 min on ice as in ref. 37. Cell lysates were then tested as in ref. 38. Briefly, cells were resuspended in 100 µl of lysis buffer (0.1 M NaOH, 0.05 M EDTA, 2% SDS, 2% β-mercaptoethanol, one complete protease inhibitor cocktail tablet (PiC, Roche) per 50 ml, one phosphatase inhibitor cocktail tablet (PhosSTOP, Roche) per 10 ml). For the detection of poly-ubiquitinated proteins, the lysis buffer is supplemented with 5 mM *N*-ethylmaleimide (Sigma-Aldrich). Lysates were incubated at 90 °C for 10 min. 2.5 µl of 4 M acetic acid were subsequently added before vortexing for 30 s. Lysates were incubated at 90 °C for 10 min and then cleared by centrifugation for 10 min at 16,000*g*. Supernatants were transferred to a clean tube and protein concentrations were measured by monitoring the OD_{280}. Protein concentrations were equilibrated to 1 µg of total proteins per µl, and 80 µl of lysates were mixed with 20 µl of 5× loading buffer (0.25 M Tris-HCl (pH 6.8), 10% SDS, 50% glycerol, 0.05% bromophenol blue). 15 µg of total protein extract was loaded on Bolt 4–12% Bis-Tris Plus gels (Life Technologies) and resolved in MES buffer. Gel-separated protein samples were transferred to nitrocellulose membranes (Life Technologies). Membranes were cut and their fragments were incubated with antibodies to Kar2 (sc-33630; Santa Cruz Biotechnology, 1:1,000), GFP (ab290; Abcam, 1:5,000), HA (mHA.11; Covance, 1:2,000), TAP (CAB1001; Pierce, 1:1,000), ubiquitin (646302 (P4D1); BioLegend, 1:1,000), Adc17 (Bertolotti laboratory[10]; 1:1,000), P-T737-Sch9 (Maeda laboratory[16]; 1:5,000), Mpk1 (sc-6803; Santa Cruz Biotechnology, 1:1,000), Hog1 (sc-9079; Santa Cruz Biotechnology, 1:1,000), P-Mpk1 (catalogue number 9101; Cell Signaling Technology, 1:1,000), Rpt5 (BML-PW8245; Enzo life sciences, 1:5,000), 20S core subunits (CP) (BML-PW9355; Enzo life sciences, 1:2,000), Nas6 (ab91447; Abcam, 1:1,000) and Nas2, Hsm3 and Rpn14 (Hochstrasser laboratory[7]; 1:1,000). Proteins were visualized by ECL Prime (GE Healthcare) using chemi-Smart 5000 or ChemiDoc Touch equipments (Bio-Rad).

For analyses of the phosphorylation status of Sch9, cell aliquots were taken at the indicated times and mixed with trichloroacetic acid (TCA) at a final concentration of 6%. Cell lysates were then prepared as described previously[39].

Native PAGE in yeast. 30 ml of exponentially growing cells adjusted to an OD_{600} of 0.2 were treated with 5 µg ml^{-1} tunicamycin, 0.2 µg ml^{-1} rapamycin or DMSO for 3 h at 32 °C. Cells were then harvested, washed in ice-cold water, resuspended in native lysis buffer (50 mM Tris-HCl (pH 7.4), 1 mM EDTA, 5 mM $MgCl_2$, 1 mM DTT, 2 mM ATP) as in ref. 40, and disrupted with glass beads (10 times for 30 s) at 4 °C. After removal of the glass beads, the extracts were cleared by centrifugation at 14,500*g* for 10 min at 4 °C. Protein concentration was measured by monitoring OD_{280} and 80 µl of adjusted extracts were mixed with 20 µl of 5× native loading buffer (0.25 M Tris-HCl (pH 6.8), 50% glycerol, 0.05% bromophenol blue). 25 µg of each extract were subjected to 4.2% native PAGE. In-gel peptidase assay was performed as described previously[10] before being transferred to nitrocellulose membranes. Membranes were incubated with antibodies to 20S (PW9355; Biomol, 1:2,000) and Rpt5 (PW8245; Biomol, 1:1,000). Proteins were visualized by ECL Prime (GE Healthcare).

Microscopy. Images of yeast cells carrying a GFP-tagged SFP1 at the endogenous locus were taken using Zeiss-710 confocal microscope. The excitation laser wavelength, emission detection bands and pinhole diameter were chosen based on the manufacturer's recommended settings for Hoechst 33342 and GFP. The laser power and detector gain settings were adjusted to avoid saturation.

Quantitative RT–PCR. Total yeast RNA was extracted as previously described[41]. 15 µg of purified RNA was treated with the Turbo DNase kit (Ambion) and 1 µg of DNA-free RNA was synthesized into cDNA using the iScript cDNA synthesis kit (Bio-Rad laboratories). cDNA was diluted 1:10 before the quantitative RT–PCR was performed.

Quantitative RT–PCR with primers *alg9* (forward): cacggatagtggctttggtgaacaattac, *alg9* (reverse): tatgattatctggcagcaggaaagaacttggg, *rpl18a* (forward): gtgccagagccaagattgtt, *rpl18a* (reverse): tggagctctgacagctaattga, *pre4* (forward): tgaaaatgcgtatgacaatcct, *pre4* (reverse): tcaaaaatatagctgggttcgag, *pre10* (forward): aagtggctcttattggggcta, *pre10* (reverse): ttcgcagattgcctaccttt, *rpt5* (forward): gcaaagaaccatgctggaat, *rpt5* (reverse): tgacacgatcatcggagcta, *rpt6* (forward): ttccattggctctactcgtg, *rpt6* (reverse): aaacccgtccaattggttta, *adc17* (forward): cgacgacttggagaacattg, *adc17* (reverse): caatgcgtccactctctcat, *nas6* (forward): tccaaaccttccttgttgcta, *nas6* (reverse): tgcttggaaagaaactgacca, *nas2* (forward): ctagaggcgtatttcagtgtgc, *nas2* (reverse): tcaccaacgcagagtccat, *hsm3* (forward): aaaatttctgctcaatgagatgc, *hsm3* (reverse): gcgctcccatcacctatc, *rpn14* (forward): tgccataatagaccgaggaag, *rpn14* (reverse): aggcgaattgtaccatccaa was performed using SYBR Select Master Mix (4472908; Applied Biosystems) on a ViiA 7 system (Life technologies). Expression of each gene was normalized to the housekeeping gene *ALG9* and expressed as fold change after 2 h rapamycin treatment calculated using Paffl equation.

Mammalian cell culture. HeLa cells were from IGBMC (Strasbourg, France) with authentication and they were not used beyond passage 20 from original derivation. HeLa cells were routinely tested for mycoplasma contaminations. HeLa cells were cultured in minimum essential media (MEM) (11095-080; Life Technologies) supplemented with L-glutamine-penicillin-streptomycin solution (G6784; Sigma-Aldrich) and containing 10% fetal bovine serum (FBS). The medium was changed every 24 h. Medium replenishment experiment was carried out using DMEM (11960-044; Life technologies, (high glucose, no glutamine)) supplemented with L-glutamine-penicillin-streptomycin solution (G6784; Sigma-Aldrich) and containing 10% FBS.

Mammalian cell treatments. For mTOR inhibition by Torin-1, cells were plated in 6-well plates at a density of 400,000 cells per well. The medium was changed 24 h after plating and a final concentration of 250 nM Torin-1, 200 nM rapamycin or DMSO was directly added to the medium 48 h after plating (confluence: 85–95%) for the indicated time. For starvation experiments, cells were plated in 6-well plates at a density of 400,000 cells per well. The medium was changed 24 h after plating. 48 h after plating, HeLa cells were washed twice with PBS before being cultured in Earle's Balanced Salt Solution (EBSS) for the indicated time points. For medium replenishment experiments, cells were plated in 6-well plates at a density of 400,000 cells per well. The medium was changed 24 h after plating. 48 h after plating, HeLa cells were washed twice with PBS before being cultured in fresh DMEM for the indicated time points.

Immunoblot analyses in mammalian cells. Cells were rinsed twice with ice-cold PBS, harvested by centrifugation and lysed in 100 µl of ice-cold lysis buffer (50 mM Tris-HCl (pH 7.4), 150 mM NaCl, 1% Triton X-100, 0.1% SDS, 1% sodium deoxycholate, one complete protease inhibitor cocktail tablet (PiC, Roche) per 50 ml, one phosphatase inhibitor cocktail tablet (PhosSTOP, Roche) per 10 ml). Lysates were then sonicated for 3 min (1 s on/1 s off). The soluble fractions from cell lysates were isolated by centrifugation at 16,000*g* for 10 min at 4 °C and protein concentrations were measured using BCA protein assay kit (Thermo scientific) and adjusted to 1 µg of total proteins per µl. 80 µl of adjusted protein extracts were mixed with 20 µl of 5× loading buffer (0.25 M Tris-HCl (pH 6.8), 10% SDS, 50% glycerol, 0.05% bromophenol blue). 15 µg of total protein extract was loaded on Bolt 4–12% Bis-Tris Plus gels (Life Technologies) and resolved in MES buffer. Gel-separated protein samples were transferred to nitrocellulose membranes (Life Technologies). Membranes were cut and their fragments were incubated with antibodies to P-p70-S6 Kinase (P-S6K1) (catalogue number 9205; Cell Signaling Technology, 1:1,000), p70-S6 kinase (S6K1) (catalogue number 92vh02; Cell Signaling Technology, 1:1,000), Rpt6 (SUG-1B8; Euromedex, 1:5,000), Alpha-7 (PW8110; Biomol, 1:1,000), p27 (Psmd9) (WH0005715M1; Sigma-Aldrich, 1:1,000), p28 (Psmd10) (catalogue number 12985; Cell Signaling Technology, 1:1,000), S5b (Psmd5) (LS-C133418; LifeSpan BioSciences inc, 1:1,000), Rpn14 (Paaf1) (ab103566; Abcam, 1:1,000), actin (ab3280; Abcam, 1:1,000), ERK5 (E1523, Sigma-Aldrich, 1:1,000) and POMP (ab170865; Abcam, 1:1,000). Proteins were visualized by ECL Prime (GE Healthcare) using chemi-Smart 5000 or ChemiDoc Touch equipments (Bio-Rad).

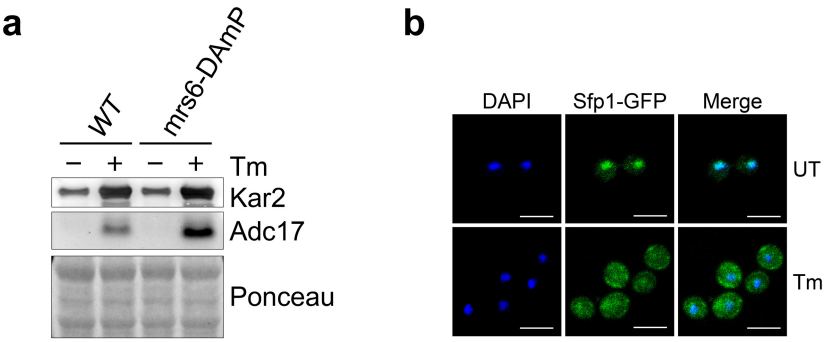
For native PAGE, cells were rinsed twice with ice-cold PBS, harvested by centrifugation and lysed in 200 µl of native lysis buffer (50 mM Tris-HCl (pH 7.4), 1 mM EDTA, 5 mM $MgCl_2$, 1 mM DTT, 2 mM ATP) as in ref. 40 and disrupted with glass beads (3 times for 20 s) at 4 °C. After removal of the glass beads, the extracts were cleared by centrifugation at 14,500*g* for 10 min at 4 °C. Protein concentration was measured by monitoring OD_{280} and 80 µl of adjusted extracts were mixed with

20 μl of 5× native loading buffer (0.25 M Tris-HCl (pH 6.8), 50% glycerol, 0.05% bromophenol blue). 25 μg of each extract were subjected to 4.2% native PAGE. In-gel peptidase assay was performed as described previously[10] before the samples were transferred to nitrocellulose membranes. Membranes were incubated with antibodies to Alpha7 (PW8110; Biomol, 1:1,000) and Rpt6 (SUG-1B8; Euromedex, 1:5,000). Proteins were visualized by ECL Prime (GE Healthcare).

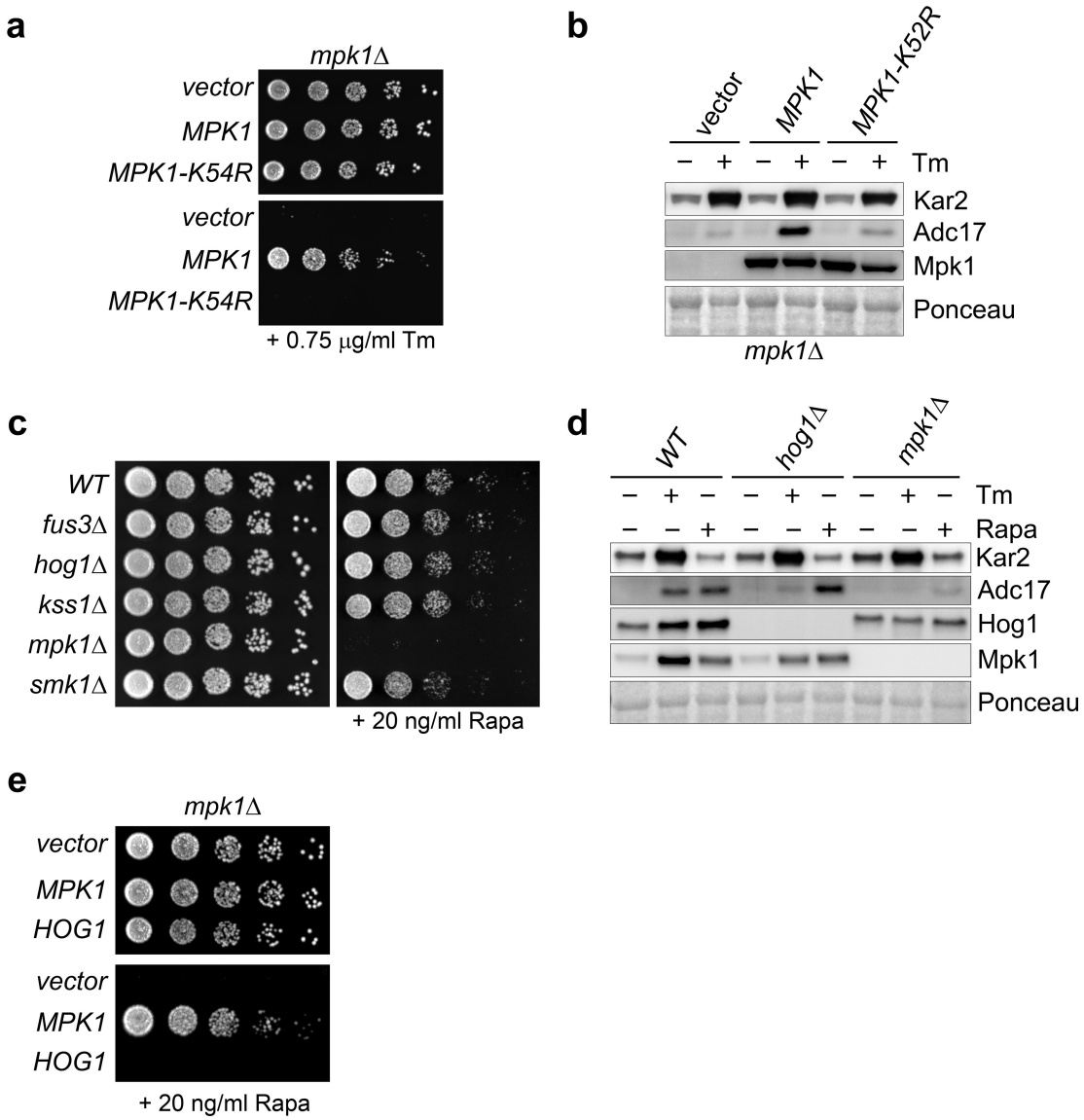
RNA interference. ON-TARGET plus SMARTpool siRNA for ERK5 or non-targeting control (Dharmacon) were used in knockdown experiments. HeLa cells (200,000 cells per well) were plated in 6-well plates. 24 h after plating, media were replenished and siRNAs were delivered into cells using RNAiMAX (catalogue number 13778075 from Invitrogen) according to the manufacturer's instructions. The medium was changed every 24 h post-transfection for a total of 3 days. Cells were then harvested and analysed by immunoblot.

Statistical analysis. Representative results of at least three independent experiments (biological replicates) are shown in all panels. GraphPad Prism software was used for all statistical analyses. Data are presented as mean and standard deviations. For immunoblot quantifications, level of each protein was normalized to PGK1 in yeast and β-actin in mammalian cells and expressed as fold change. Data were analysed using unpaired Student's *t*-test or repeated measures analysis of variance (one-way ANOVA or two-way ANOVA where indicated). The level of significance was set at $*P \leq 0.05$; $**P \leq 0.01$; $***P \leq 0.001$; NS, not significant.

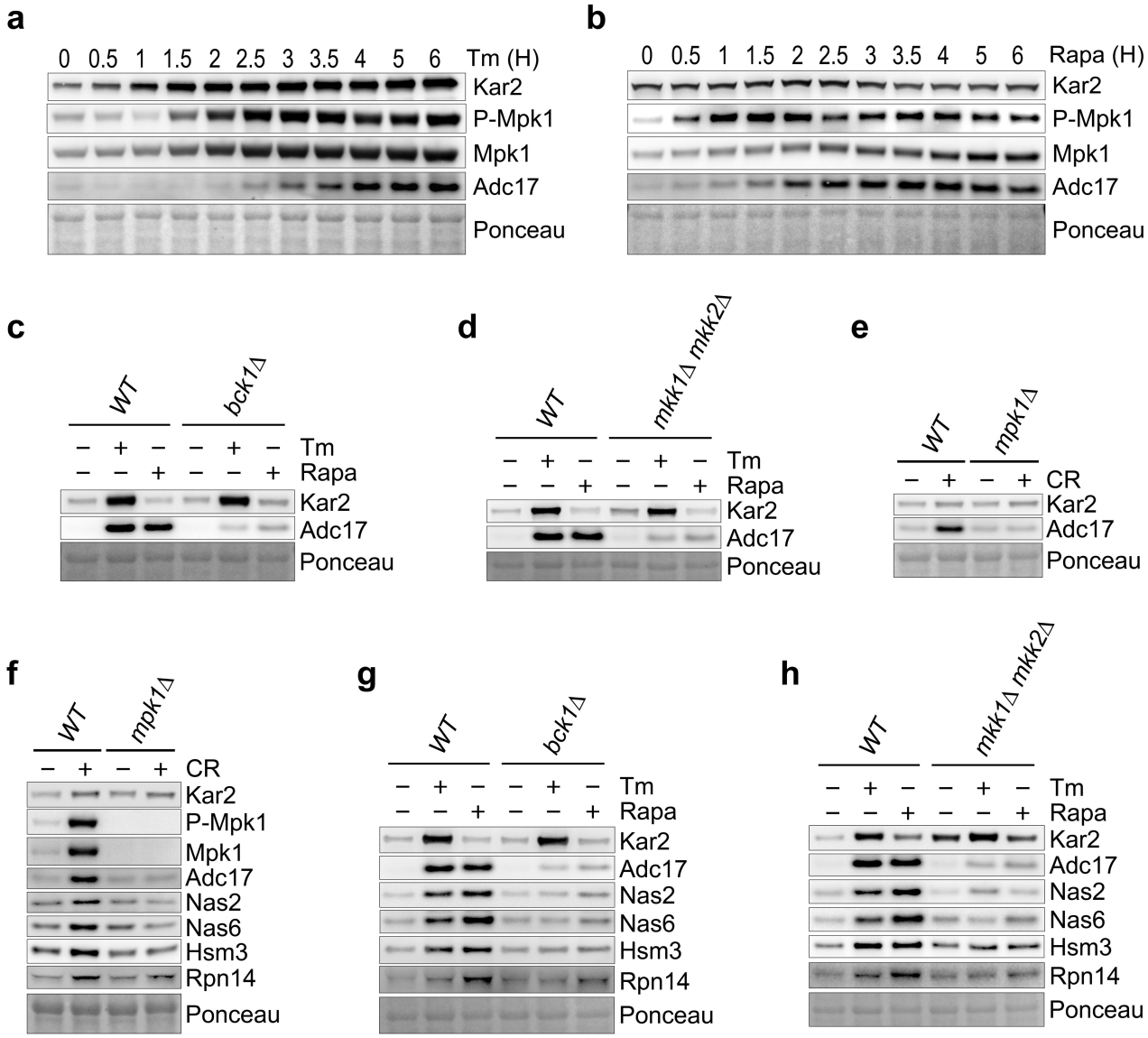
36. Gietz, R. D. & Woods, R. A. Yeast transformation by the LiAc/SS carrier DNA/PEG method. *Methods Mol. Biol.* **313,** 107–120 (2006).
37. Zhang, T. *et al.* An improved method for whole protein extraction from yeast *Saccharomyces cerevisiae*. *Yeast* **28,** 795–798 (2011).
38. von der Haar, T. Optimized protein extraction for quantitative proteomics of yeasts. *PLoS One* **2,** e1078 (2007).
39. Urban, J. *et al.* Sch9 is a major target of TORC1 in *Saccharomyces cerevisiae*. *Mol. Cell* **26,** 663–674 (2007).
40. Elsasser, S., Schmidt, M. & Finley, D. Characterization of the proteasome using native gel electrophoresis. *Methods Enzymol.* **398,** 353–363 (2005).
41. Knutson, B. A. & Hahn, S. Domains of Tra1 important for activator recruitment and transcription coactivator functions of SAGA and NuA4 complexes. *Mol. Cell. Biol.* **31,** 818–831 (2011).



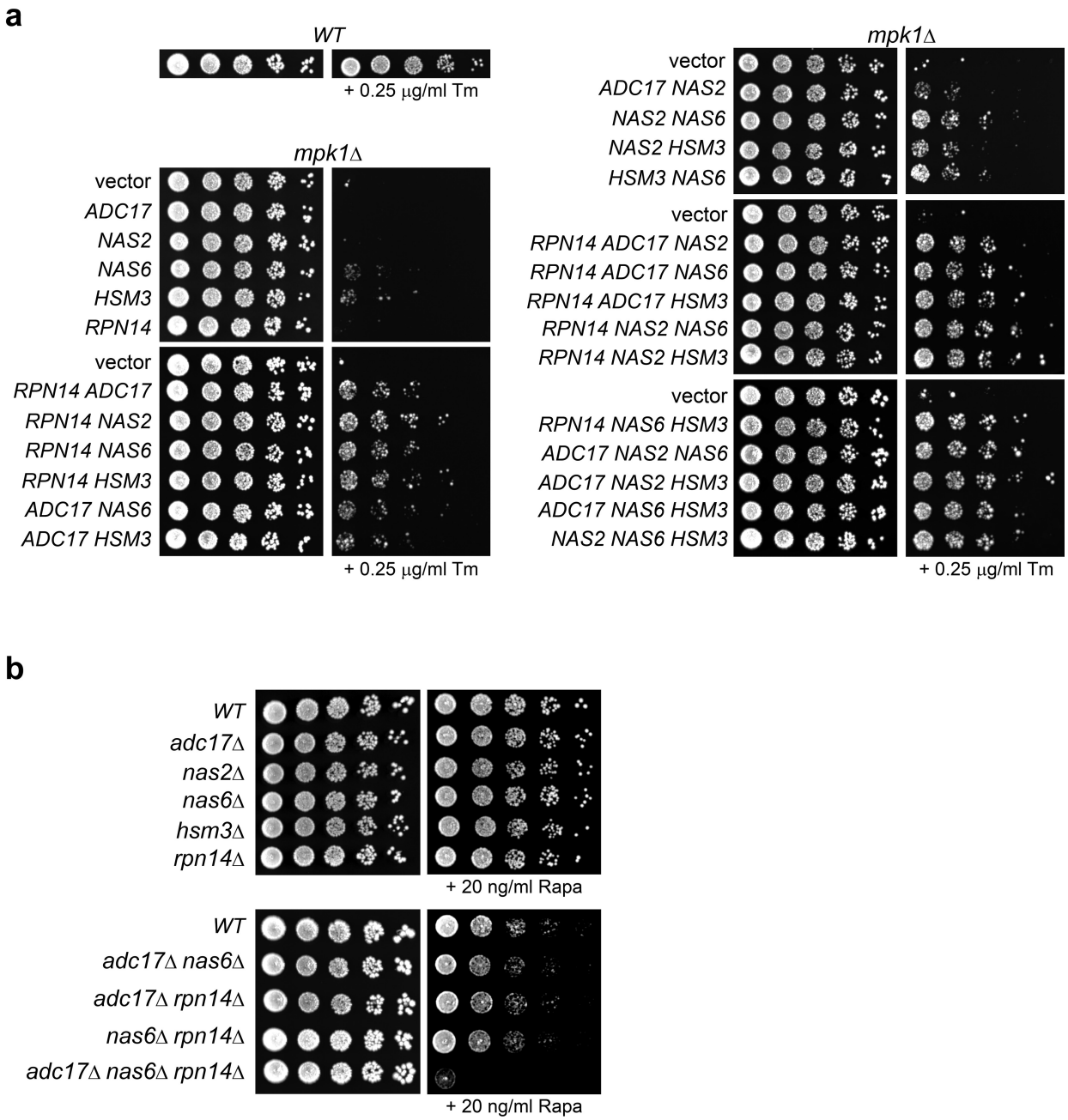
Extended Data Figure 1 | Adc17 induction is increased in mrs6-DAmp cells and occurs when Sfp1 is cytosolic. a, Immunoblots of the indicated proteins in lysates of wild-type and Mrs6-hypomorphic (mrs6-DAmP) yeast strains ± tunicamycin for 4 h. **b**, Representative images of yeast cells carrying a GFP-tagged SFP1 at the endogenous locus, ± tunicamycin for 4 h. Scale bar, 5 μm. Representative results of at least three independent experiments (biological replicates) are shown.



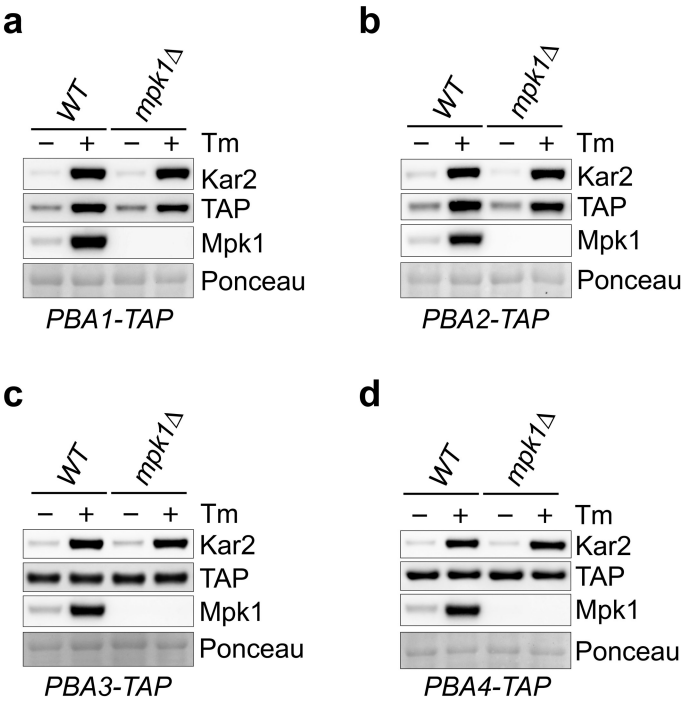
Extended Data Figure 2 | Mpk1 is essential for tunicamycin and rapamycin survival and Adc17 induction. a, *mpk1Δ* cells transformed with wild-type *MPK1* or a kinase-dead allele (*MPK1-K52R*) or empty vector were spotted in a sixfold dilution and grown on plates containing or lacking tunicamycin. **b**, Immunoblots of lysates of yeast strains shown in **a**, cultured for 4 h ± tunicamycin. **c**, Cells of the indicated genotype were spotted in a sixfold dilution and grown for 3 days at 30 °C on plates containing or lacking rapamycin. **d**, Immunoblots of lysates from wild-type and MAPK genetic deletion mutant yeast cells cultured for 4 h ± tunicamycin or rapamycin. **e**, Same as in **a**, using *mpk1Δ* cells transformed with empty vector or a vector encoding *MPK1* or *HOG1*. Representative results of at least three independent experiments (biological replicates) are shown.



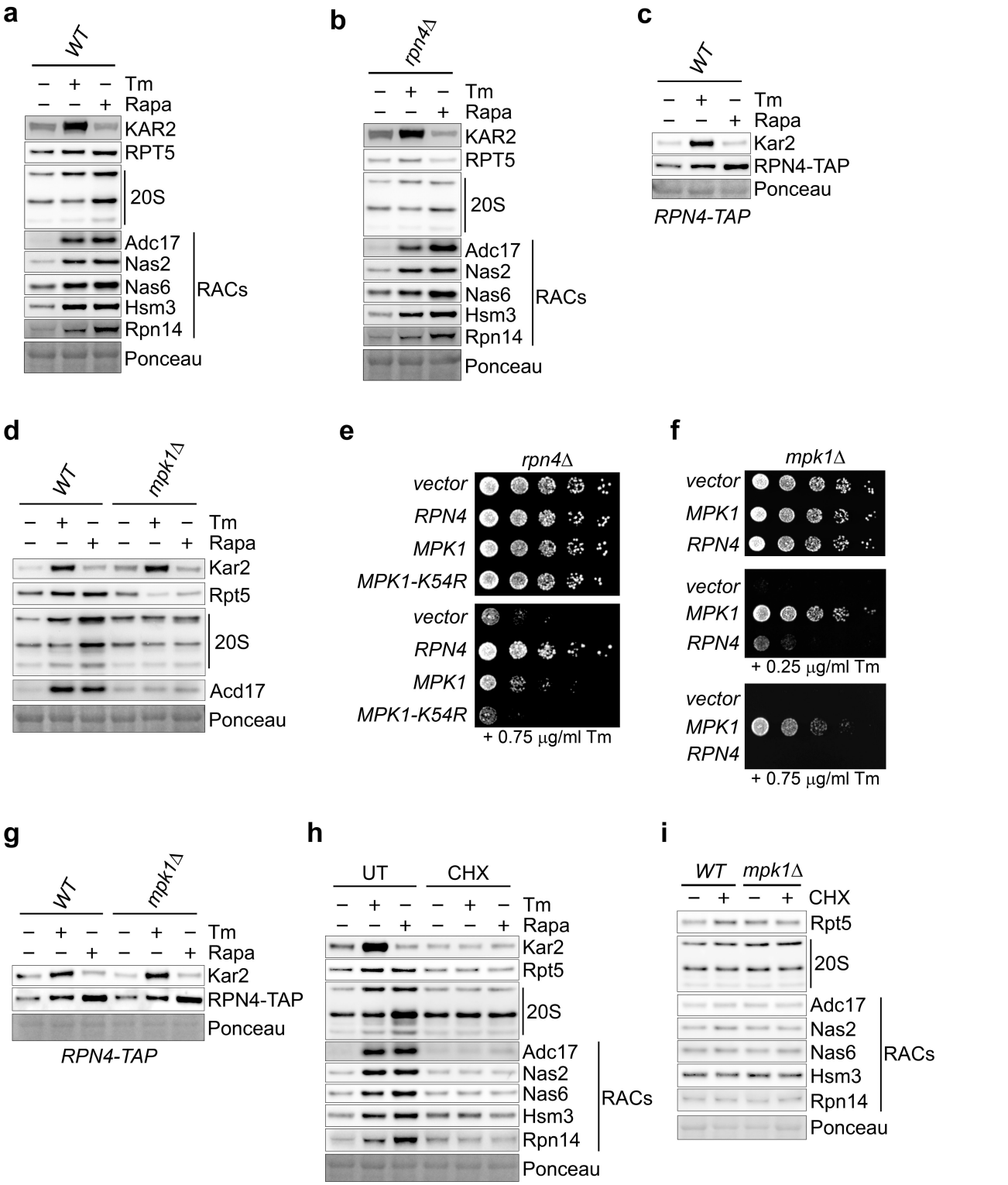
Extended Data Figure 3 | Mpk1 MAPK pathway is essential for stress-mediated RACs induction. a, **b**, Immunoblots of the indicated proteins in lysates of wild-type yeast cells ± tunicamycin (**a**) or rapamycin (**b**) for the indicated time. **c**, **g**, Immunoblots of the indicated proteins in lysates of wild-type and *bck1*Δ cells cultured ± tunicamycin or rapamycin for 4 h. **d**, **h**, Immunoblots of the indicated proteins in lysates of wild-type and *mkk1*/2Δ cells cultured ± tunicamycin or rapamycin for 4 h. **e**, **f**, Immunoblots of the indicated proteins in lysates of wild-type or *mpk1*Δ cells ± 50 μg ml^{-1} Congo red (CR) for 4 h. Representative results of at least three independent experiments (biological replicates) are shown.



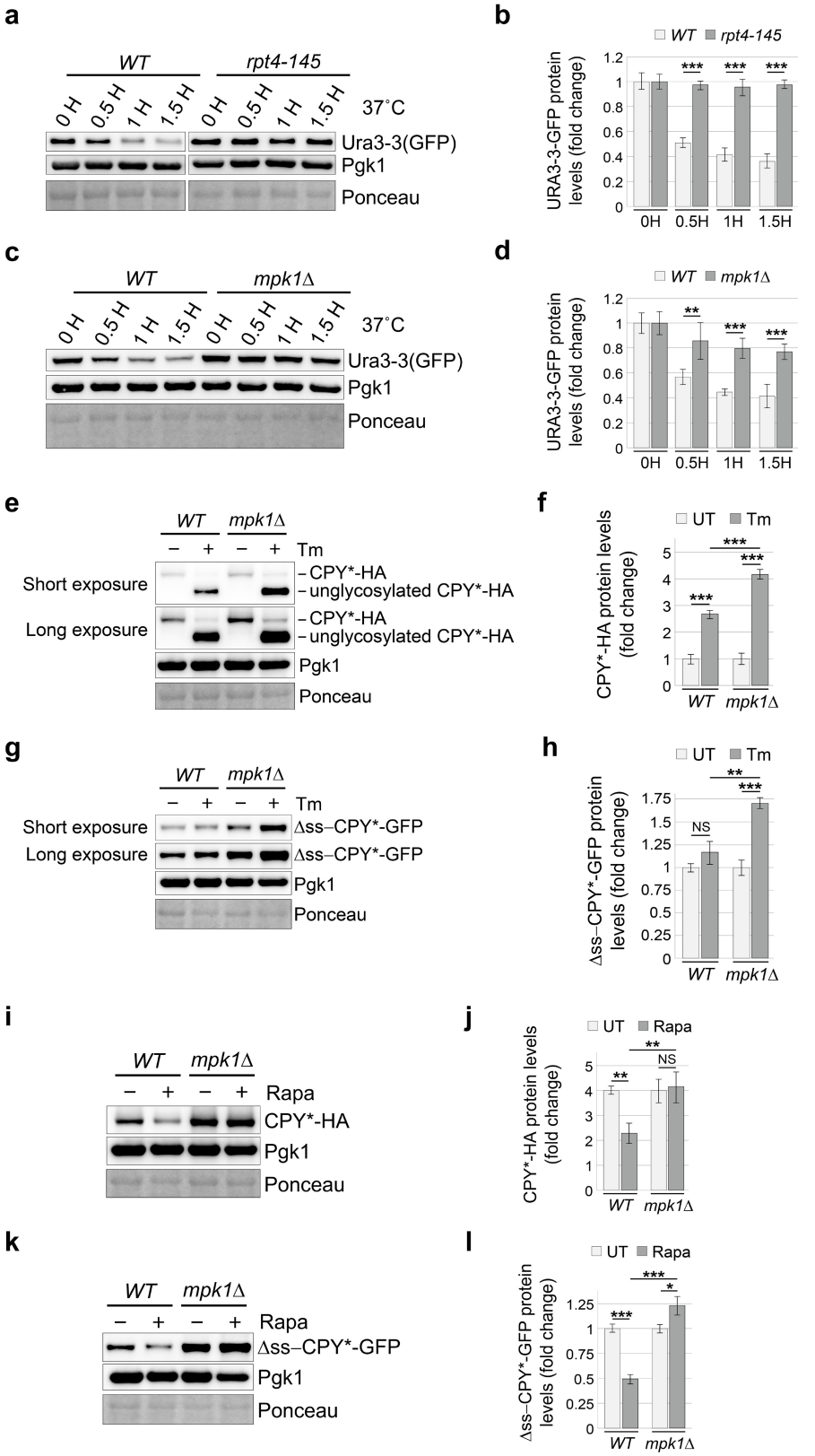
Extended Data Figure 4 | Induction of RACs under challenging conditions is an important function of Mpk1. a, Wild-type cells or *mpk1Δ* cells transformed with one or combinations of two or three RACs were spotted in a sixfold dilution and grown on plates containing or lacking tunicamycin, where indicated. **b**, Multiple-deletion yeast strains of different RACs were spotted in a sixfold dilution and grown for 3 days at 33 °C on plates containing or lacking rapamycin. Representative results of at least three independent experiments (biological replicates) are shown.



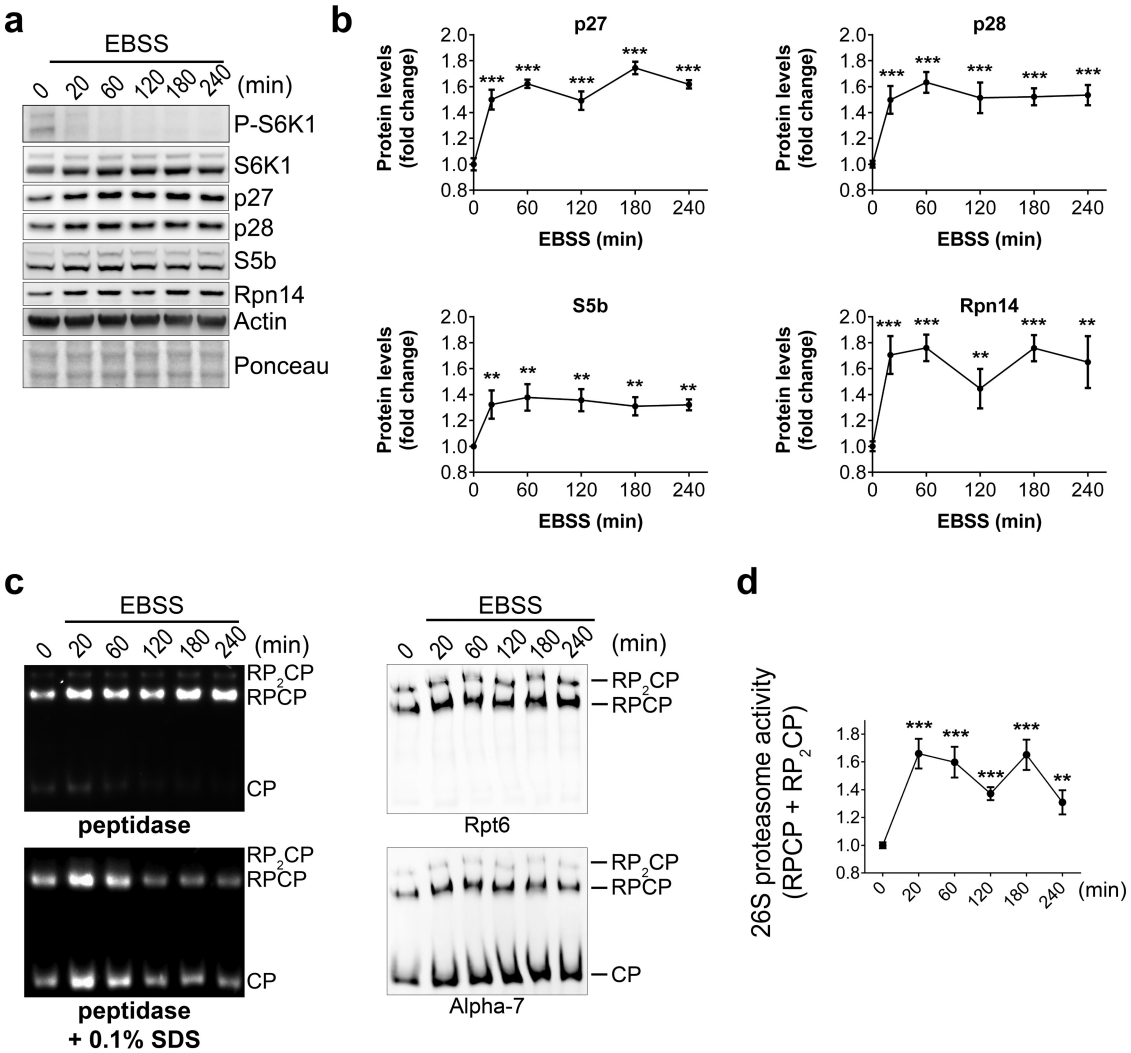
Extended Data Figure 5 | Pba1 and Pba2 are induced by tunicamycin in a Mpk1-independent manner. a–d, Immunoblots of the indicated proteins in lysates of wild-type yeast cells carrying a TAP-tagged Pba1 (**a**), Pba2 (**b**), Pba3 (**c**) and Pba4 (**d**) at the endogenous locus ± tunicamycin for 3 h. Representative results of at least three independent experiments (biological replicates) are shown.



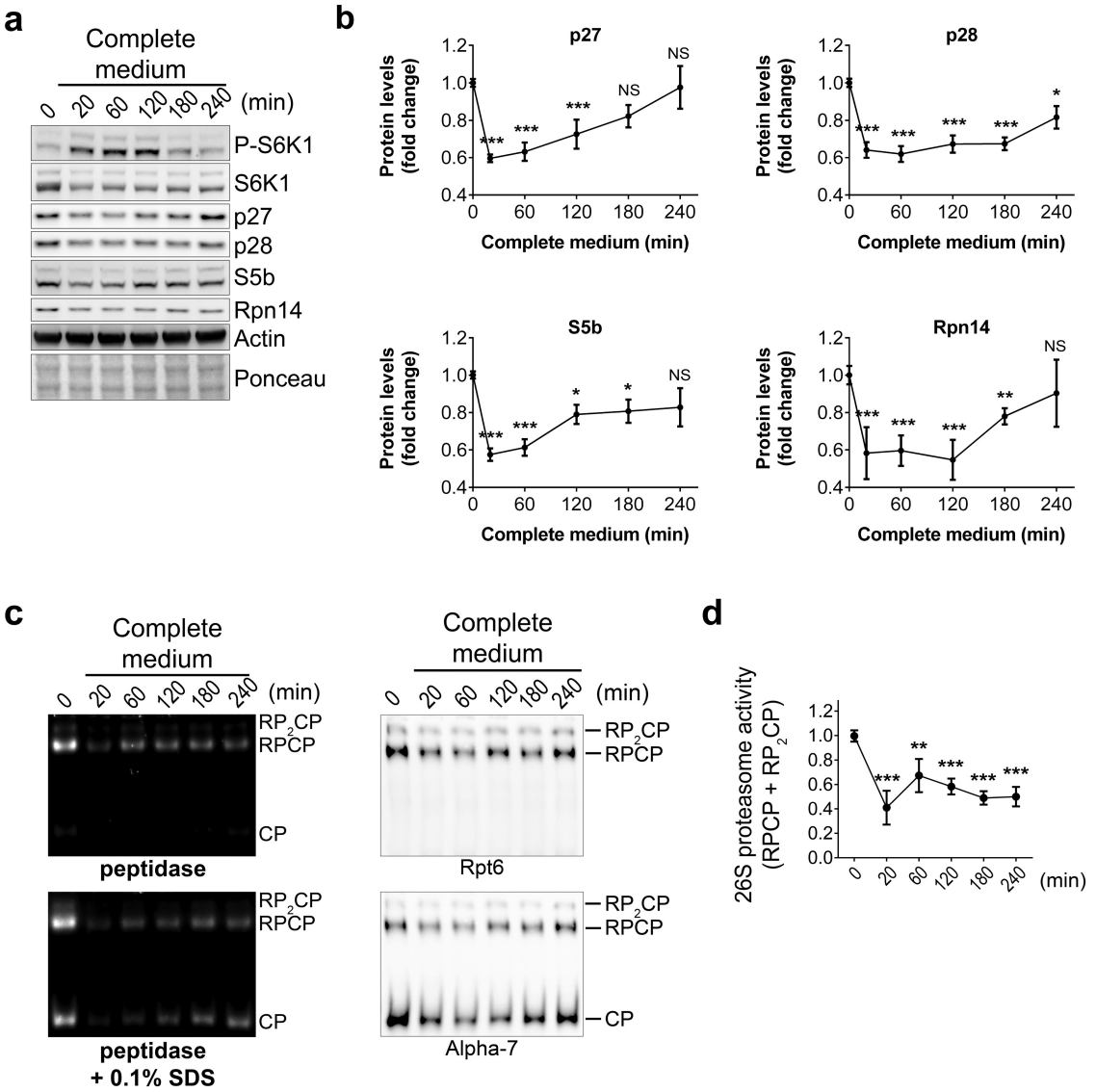
Extended Data Figure 6 | Mpk1 post-transcriptionally regulates proteasome subunits and RACs. a, **b**, Immunoblots of the indicated proteins in lysates of wild-type (**a**) and *rpn4Δ* (**b**) cells ± tunicamycin or rapamycin for 4 h. **c**, Immunoblots of the indicated proteins in lysates of wild-type yeast cells carrying a TAP-tagged RPN4 at the endogenous locus ± tunicamycin or rapamycin for 4 h. **d**, Immunoblots of the indicated proteins in lysates of wild-type and *mpk1Δ* cells ± tunicamycin or rapamycin for 4 h. **e**, *rpn4Δ* cells transformed with *RPN4*, *MPK1*, a kinase-dead allele of *MPK1* (*MPK1-K52R*) or empty vector were spotted in a sixfold dilution and grown on plates containing or lacking tunicamycin. **f**, *mpk1Δ* cells transformed with *MPK1*, *RPN4* or empty vector were spotted in a sixfold dilution and grown on plates containing or lacking tunicamycin where indicated. **g**, Immunoblots of the indicated proteins in lysates of wild-type and *mpk1Δ* cells carrying a TAP-tagged RPN4 at the endogenous locus ± tunicamycin or rapamycin for 4 h. **h**, **i**, Immunoblots of the indicated proteins in lysates of wild-type (**h**, **i**) and *mpk1Δ* (**i**) cells treated with different combinations of drugs: 5 μg ml^{-1} tunicamycin, 0.2 μg ml^{-1} rapamycin and 35 μg ml^{-1} cycloheximide, where indicated for 4 h. Representative results of at least three independent experiments (biological replicates) are shown.



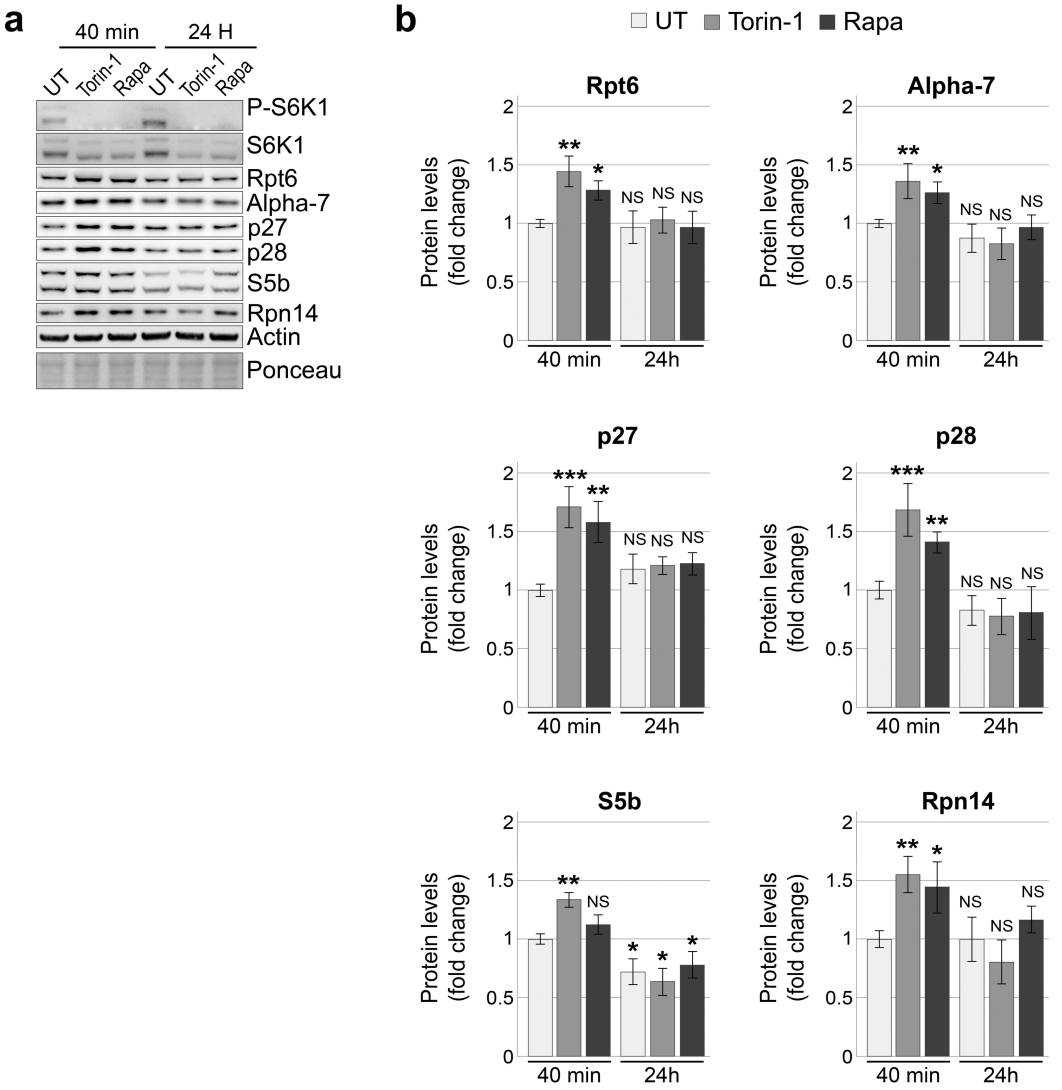
Extended Data Figure 7 | Mpk1 maintains the adequate levels of proteasome required to sustain protein degradation. a, **c**, Yeast cells of the indicated genotype expressing GFP-tagged Ura3-3 proteins were treated with cycloheximide and incubated at 37 °C for the indicated time. **b**, **d**, Quantifications from three independent experiments (biological replicates) such as the one shown in **a** and **c**. **e**, **g**, Cells of the indicated genotype expressing CPY*–HA (**e**) or Δss-CPY*–GFP (**g**) proteins were treated with tunicamycin for 4 h. **f**, **h**, Quantifications from three independent experiments (biological replicates) such as the one shown in **e** and **g**. **i**, **k**, Cells of the indicated genotype expressing CPY*–HA (**i**) or Δss-CPY*–GFP (**k**) proteins were treated with rapamycin for 4 h. **j**, **l**, Quantifications from three independent experiments (biological replicates) such as the one shown in **i** and **k**. **b**, **d**, **f**, **h**, **j** and **l**, Data are mean ± s.d. $n = 3$ biological replicates. $^{*}P \leq 0.05$; $^{**}P \leq 0.01$; $^{***}P \leq 0.001$; NS, not significant (two-way ANOVA).



Extended Data Figure 8 | Starvation inhibits TORC1 signalling, induces mammalian RACs and increases proteasome abundance. **a**, **b**, Immunoblots (**a**) and quantification (**b**) of the indicated proteins in lysates of HeLa cells after EBSS (Earle's balanced salt solution) treatment for the indicated time. **c**, HeLa cell extracts following EBSS treatment for the indicated time were resolved on native PAGE (4.2%) and monitored using the fluorogenic substrate Suc-LLVY–AMC or by immunoblots. **d**, Quantification of the 26S proteasome activity (RPCP and RP_2CP) of experiments such as the one shown in **c**. **b**, **d**, Data are mean ± s.d. $n = 3$ biological replicates. $^*P \leq 0.05$; $^{**}P \leq 0.01$; $^{***}P \leq 0.001$; NS, not significant (one-way ANOVA). Representative results of at least three independent experiments (biological replicates) are shown.



Extended Data Figure 9 | TORC1 activation by nutrient replenishment decreases the abundance of RACs as well as 26S proteasome. **a**, **b**, Immunoblots (**a**) and quantification (**b**) of the indicated proteins in lysates of HeLa cells after replenishment with rich complete medium for the indicated time. **c**, Native PAGE (4.2%) of cell extracts from HeLa cells following media replenishment as in **a**, monitored by the fluorogenic substrate Suc-LLVY–AMC or by immunoblots. **d**, Quantification of the 26S proteasome activity (RPCP and RP_2CP) of experiments such as the one shown in **c**. **b**, **d**, Data are mean ± s.d. $n = 3$ biological replicates. $^*P \leq 0.05$; $^{**}P \leq 0.01$; $^{***}P \leq 0.001$; NS, not significant (one-way ANOVA). Representative results of at least three independent experiments (biological replicates) are shown.



Extended Data Figure 10 | mTORC1 inhibition by Torin-1 and rapamycin acutely induced the RACs. a, **b**, Immunoblots (**a**) and quantification (**b**) of the indicated proteins in lysates of HeLa cells treated with 250 nM Torin-1 or 200 nM rapamycin for the indicated time. Data are mean ± s.d. $n = 3$ biological replicates. $^{*}P \leq 0.05$; $^{**}P \leq 0.01$; $^{***}P \leq 0.001$; NS, not significant (two-way ANOVA). Representative results of at least three independent experiments (biological replicates) are shown.

LETTER

doi:10.1038/nature18940

Heating of Jupiter's upper atmosphere above the Great Red Spot

J. O'Donoghue[1], L. Moore[1], T. S. Stallard[2] & H. Melin[2]

The temperatures of giant-planet upper atmospheres at mid- to low latitudes are measured to be hundreds of degrees warmer than simulations based on solar heating alone can explain[1–4]. Modelling studies that focus on additional sources of heating have been unable to resolve this major discrepancy. Equatorward transport of energy from the hot auroral regions was expected to heat the low latitudes, but models have demonstrated that auroral energy is trapped at high latitudes, a consequence of the strong Coriolis forces on rapidly rotating planets[3–5]. Wave heating, driven from below, represents another potential source of upper-atmospheric heating, though initial calculations have proven inconclusive for Jupiter, largely owing to a lack of observational constraints on wave parameters[6,7]. Here we report that the upper atmosphere above Jupiter's Great Red Spot—the largest storm in the Solar System—is hundreds of degrees hotter than anywhere else on the planet. This hotspot, by process of elimination, must be heated from below, and this detection is therefore strong evidence for coupling between Jupiter's lower and upper atmospheres, probably the result of upwardly propagating acoustic or gravity waves.

On 4 December 2012 (Coordinated Universal Time, UTC) we observed Jupiter for 9 h using the SpeX spectrometer[8] on the NASA Infrared Telescope Facility. The spectrometer slit was aligned along the rotational axis in the north–south direction at local noon on the planet. This arrangement is illustrated in Fig. 1a, which contains a slit-jaw image showing bright auroral emissions at the poles as well as a localized Great Red Spot (GRS) emission enhancement at mid-latitudes. Exposures from the instrument in this setup give wavelength and intensity information as a function of latitude as shown in Fig. 1b. By exposing continuously throughout the night, we obtained longitudinal information for most of the planet (a Jovian day is 9 h 56 min long).

The spectrum in Fig. 1b shows strong emission features at six wavelengths, which appear prominently in the auroral regions and wane towards the equator. These are discrete ro-vibrational emission lines from H_3^+, a major ion in Jupiter's ionosphere, the charged (plasma) component of the upper atmosphere. The colour contours highlight the weaker emissions from this ion across the body of the planet. Far from a uniform intensity at low latitudes, there is a substantial intensity enhancement in all of the emission lines within the −13° to −27° planetocentric latitude range occupied by the GRS[9]. As seen in the coloured contours of Fig. 1b, the H_3^+ emissions are isolated in wavelength, indicating that there is no continuum reflection of sunlight at the latitudes of the GRS.

The ratio between two or more emission lines can be used to derive the temperature of the emitting ions[10,11]. With the observing geometry used here, such temperatures are altitudinally averaged 'column temperatures' of H_3^+, where the majority of H_3^+ at Jupiter has been observed to be located at altitudes between 600 km and 1,000 km above the 1-bar pressure level[12]. H_3^+ has been demonstrated to be in quasi-local thermodynamic equilibrium throughout the majority of Jupiter's upper atmosphere, meaning that derived temperatures are representative of the co-located ionosphere and (the mostly H_2) thermosphere[13]. In the Methods section we detail the data reduction techniques and temperature model fitting procedures, and in Fig. 2 we show two example model fits; only the strongest, outermost lines are used to fit temperatures, because the central H_3^+ lines are contaminated by telluric absorption. Note that, even though the H_3^+ peak intensities at the GRS (Fig. 2a) are lower than those at 45° latitude, this is a result of lower column-integrated H_3^+ densities at lower latitude. Derived temperatures remain unaffected by the density differences because they are based entirely on H_3^+ line ratios.

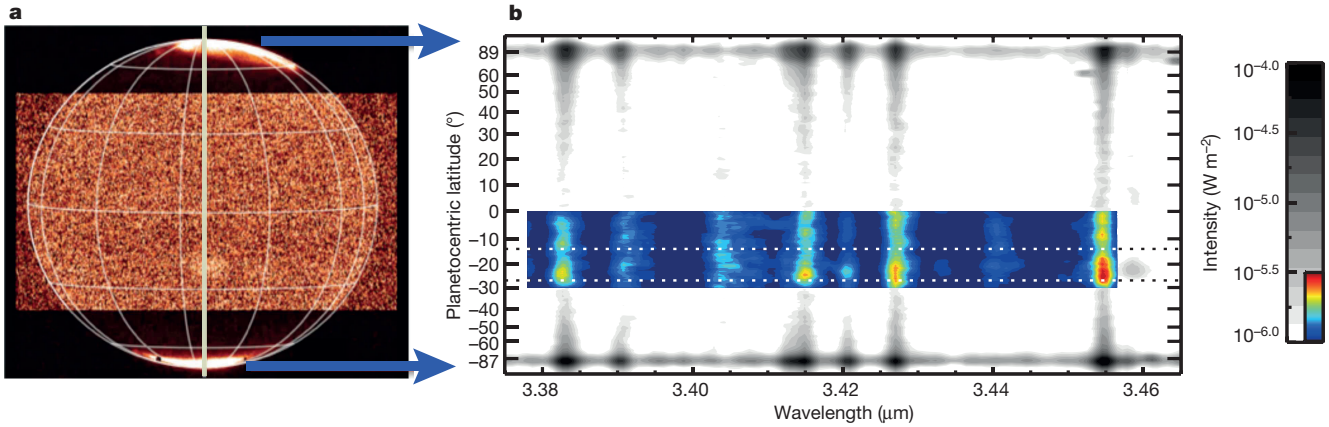


Figure 1 | The acquisition of Jovian spectra. a, Jupiter as observed by the SpeX slit-jaw imager and L-filter (3.13–3.53 μm), on 4 December 2012. Bright regions at the poles result from auroral emissions; the contrast at low and mid-latitudes has been enhanced for visibility. The vertical beige line in the middle of the image indicates the position of the spectrometer slit, which was aligned along the rotational axis. **b**, The co-added spectrum of seven GRS-containing exposures; dotted horizontal lines indicate the latitudinal range of the GRS. Further details are given in the Methods section.

[1]Center for Space Physics, Boston University, Boston 02215, USA. [2]Department of Physics and Astronomy, University of Leicester, Leicester, LE1 7RH, UK.

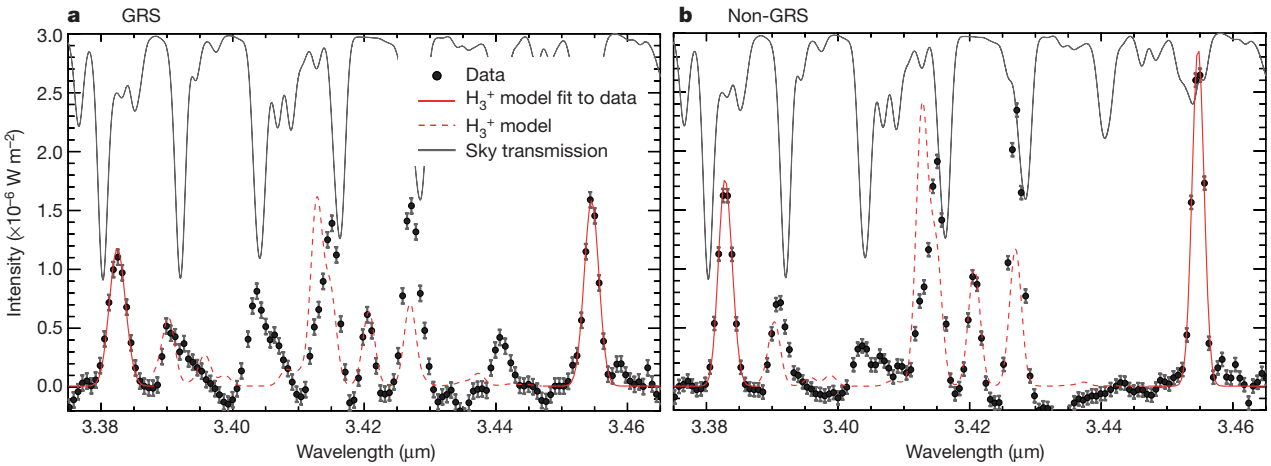


Figure 2 | Model fit to observed H_3^+ intensity as a function of wavelength. **a**, The data in Fig. 1b plotted between −13° and −19° planetocentric latitude; **b**, as for **a** but plotted between −40° and −49° planetocentric latitude. The H_3^+ model fit to the data is shown as a solid red line: only the H_3^+ lines at 3.383 μm and 3.454 μm are included in the temperature derivation (see Methods for the full list). Telluric absorption, normalized to show sky contamination, is shown in grey. The derived temperatures are 1,644 ± 161 K (**a**) and 900 ± 42 K (**b**) (±standard errors of the mean). The H_3^+ model is extended to the central region (dotted red line) based on the temperatures and densities of the fits. Intensity errors are 1σ.

The difficulty in explaining the observed upper-atmospheric temperatures of the giant planets was realized more than 40 years ago[1], and has since been termed the giant-planet "energy crisis"[2,4]. For Jupiter, only the observed temperatures within the auroral regions have been adequately explained, as the 1,000–1,400 K temperatures[14] observed there result from auroral heating mechanisms that impart 200 GW of power per hemisphere through ion-neutral collisions and Joule heating[15,16]. The low to mid-latitudes do not have such a heat source, and yet are measured to be near 800 K, which is 600 K warmer than can be accounted for by solar heating[15,17,18]. If heating does not come from above (solar heating), and cannot be produced *in situ* via magnetospheric interactions, then a solution is likely to be found below.

Gravity waves, generated in the lower atmosphere and breaking in the thermosphere, represent a potentially viable source of upper-atmospheric heating. Previous modelling studies, however, have led to inconclusive results for Jupiter: while viscous dissipation of gravity waves in Jupiter's upper atmosphere can lead to warming of the order of 10 K, sensible heat flux divergence can also lead to cooling by a similar amount, depending on the properties of the wave[6,7]. Recent re-analysis of Galileo Probe data has shown that gravity waves impart a negligible amount of heating vertically to the stratosphere (gravity-wave motion is primarily longitudinal and latitudinal) and that heating near the thermosphere is less than 1 K per Jovian day[19].

A more likely energy source is acoustic waves that heat from below (also via viscous dissipation); this form of heating requires vertical propagation of disturbances in the low-altitude atmosphere. Acoustic waves are produced above thunderstorms, and the subsequent waves have been modelled to heat the Jovian upper atmosphere by 10 K per day[20] and on Earth have been observed to heat the thermosphere over the Andes mountains[20,21]. On Jupiter, acoustic-wave heating has been modelled to potentially impart hundreds of degrees of heating to the upper atmosphere[22]. However, to the best of our knowledge, no such coupling between the lower and upper atmosphere has been directly observed for the outer planets, so vertical coupling has not been seriously considered as a solution to the giant-planet energy crisis.

Jupiter's GRS is the largest storm in the Solar System, spanning 22,000 km by 12,000 km in longitude and latitude, respectively. The GRS lies within the troposphere, with cloud tops reaching altitudes of 50 km, around 800 km below the H_3^+ layer[9]. In Fig. 3 we show (red circles) that the pattern of H_3^+ intensity seen above the GRS, when fitted to our model, gives column-averaged H_3^+ temperatures of over 1,600 K, higher than anywhere else on the planet, even in the auroral region. We also fitted temperatures to a swath of longitudes away from the GRS in order to illustrate that the enhancement in temperature occurs only within this longitude band. The latitudinal variation of temperatures away from the GRS is similar to the ranges previously observed[17], indicating that the high temperature above the GRS is localized in both latitude and longitude.

The high temperature in the northern part of the GRS provides direct observational evidence of a localized heating process. We interpret the cause of this heating to be storm-enhanced atmospheric turbulence, which arises due to the flow shear between the storm and the surrounding atmosphere. Some of these waves must then propagate vertically upwards, depositing their energy as heat through viscous dissipation. It is unknown, at present, why the two red data points at GRS latitudes (grey shaded region in Fig. 3) differ by 800 K. Perhaps there may be contamination of the H_3^+ line at 3.454 μm by the methane emission line at the same wavelength. Any additional intensity added to this H_3^+ line results in a lower temperature (for further detail see the Methods section). Thus, the temperature above the southern part

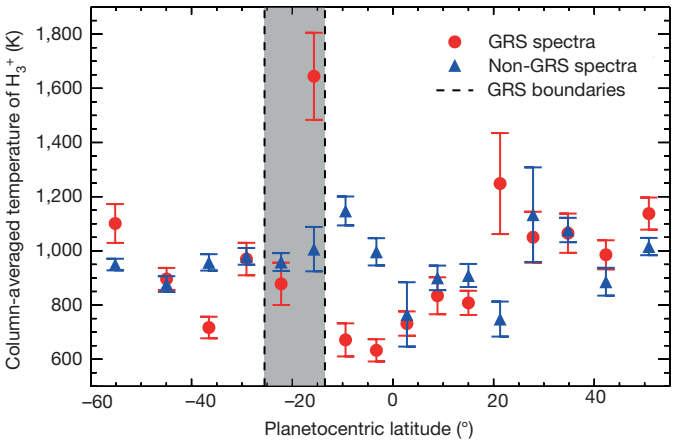


Figure 3 | Jovian H_3^+ temperatures versus planetocentric latitude. Column-averaged temperatures of H_3^+ shown here are each derived from model fits to the discrete H_3^+ emission lines as shown in Fig. 2. Red circle symbols correspond to the co-addition of GRS-related spectra (that is, from the spectral image in Fig. 1b) between 239° and 253° in Jovian system III Central Meridian Longitude (CML). The GRS latitudes are indicated by the grey shading. Blue triangle symbols were derived from exposures taken in the ranges 293°–359° and 0°–82° CML, that is, longitudes well separated from the GRS, representing the 'ordinary' background conditions based on solar heating alone. The modelled temperature of the upper atmosphere for these non-auroral regions is 203 K (ref. 1). Uncertainties are standard errors of the mean.

of the GRS may be much higher than derived, but only if methane is preferentially brighter in the south. However, as the H_3^+ and CH_4 lines at 3.454 μm are not separated spectrally in this work, it is not possible to conclude whether or not contamination is present.

An alternative physical explanation may relate to the relative velocities between the zonal wind and the GRS being greatest on the equatorward side of the storm: relative velocities are 75 m s^{-1} in the north, 15 m s^{-1} in the storm core, and 25 m s^{-1} at the poleward edge[9]. The largest relative velocities would induce the strongest flow shear, leading to the greatest turbulence and therefore the largest contribution to heating above. It is possible that evidence of such energy transfer from the lower to the upper atmosphere would be deposited en route in the intervening troposphere and upper stratosphere (0–150 km altitude), as there is a temperature enhancement of 10 K encircling the GRS at these altitudes[23,24]. However, this enhancement could also be due to the upwelling of material in the centre of the GRS, followed by increased adiabatic heating when the material downwells around the edges[24].

The only previous map of Jovian H_3^+ temperatures that contains the GRS was made using ground-based data obtained in 1993 (ref. 17). The authors of ref. 17 did not mention the GRS, as no obvious signature was present in their temperature map. However, on the basis of their temperature contours and the expected location of the GRS at the time, we estimate that there was a measured temperature enhancement of 50 K above the GRS. Such a minor temperature increase may indicate that the GRS-driven heating of Jupiter's upper atmosphere is transient, but the spatial resolution of the 1993 observations was 9,800 km per pixel (at the equator), compared with 500 km per pixel in this study. Therefore, the previous data had much cruder resolution in latitude and longitude, and any localized temperature enhancements would have been smoothed out.

In this work, the high-temperature region above the GRS is localized in latitude and longitude, indicating a large temperature gradient and perhaps a confinement by currently unknown upper-atmospheric dynamics. If wave heating driven from below is responsible for the temperatures observed in Jupiter's non-auroral upper atmosphere, then we might expect a relatively smooth temperature profile with latitude, punctuated by temperature enhancements above active storms. The GRS may then simply be the 'smoking gun' that dramatically illustrates this atmospheric coupling process, and provides the clue to solving the giant-planet energy crisis.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 14 April; accepted 4 June 2016.
Published online 27 July 2016.

1. Strobel, D. F. & Smith, G. R. On the temperature of the Jovian thermosphere. *J. Atmos. Sci.* **30,** 718–725 (1973).
2. Miller, S., Aylward, A. & Millward, G. in *The Outer Planets and their Moons* Vol. 19 of *Space Sciences Series of ISSI* **116,** 319–343 (2005).
3. Smith, C. G. A., Aylward, A. D., Millward, G. H., Miller, S. & Moore, L. E. An unexpected cooling effect in Saturn's upper atmosphere. *Nature* **445,** 399–401 (2007).
4. Yates, J. N., Achilleos, N. & Guio, P. Response of the Jovian thermosphere to a transient pulse in solar wind pressure. *Planet. Space Sci.* **91,** 27–44 (2014).
5. Smith, C. G. A. & Aylward, A. D. Coupled rotational dynamics of Jupiter's thermosphere and magnetosphere. *Ann. Geophys.* **27,** 199–230 (2009).
6. Hickey, M. P., Walterscheid, R. L. & Schubert, G. Gravity wave heating and cooling in Jupiter's thermosphere. *Icarus* **148,** 266–281 (2000).
7. Matcheva, K. I. & Strobel, D. F. Heating of Jupiter's thermosphere by dissipation of gravity waves due to molecular viscosity and heat conduction. *Icarus* **140,** 328–340 (1999).
8. Rayner, J. T. *et al.* SpeX: a medium-resolution 0.8–5.5 micron spectrograph and imager for the NASA infrared telescope facility. *Publ. Astron. Soc. Pacif.* **115,** 362–382 (2003).
9. Parisi, M., Galanti, E., Finocchiaro, S., Iess, L. & Kaspi, Y. Probing the depth of Jupiter's Great Red Spot with the Juno gravity experiment. *Icarus* **267,** 232–242 (2016).
10. Melin, H., Miller, S., Stallard, T., Smith, C. & Grodent, D. Estimated energy balance in the Jovian upper atmosphere during an auroral heating event. *Icarus* **181,** 256–265 (2006).
11. O'Donoghue, J. *et al.* Conjugate observations of Saturn's northern and southern H_3^+ aurorae. *Icarus* **229,** 214–220 (2014).
12. Uno, T. *et al.* Vertical emissivity profiles of Jupiter's northern H_3^+ and H_2 infrared auroras observed by Subaru/IRCS. *J. Geophys. Res.* **119,** 10,219–10,241 (2014).
13. Miller, S. *et al.* H_3^+: the driver of giant planet atmospheres. *Phil. Trans. R. Soc. Lond.* **364,** 3121–3137 (2006).
14. Lystrup, M. B., Miller, S., Dello Russo, N., Vervack, R. J. Jr & Stallard, T. First vertical ion density profile in Jupiter's auroral atmosphere: direct observations using the Keck II telescope. *Astrophys. J.* **677,** 790–797 (2008).
15. Yelle, R. V. & Miller, S. in *Jupiter's Thermosphere and Ionosphere* 185–218 (Cambridge Univ. Press, 2004).
16. Cowley, S. W. H. *et al.* A simple axisymmetric model of magnetosphere-ionosphere coupling currents in Jupiter's polar ionosphere. *J. Geophys. Res.* **110,** 11209 (2005).
17. Lam, H. A. *et al.* A baseline spectroscopic study of the infrared auroras of Jupiter. *Icarus* **127,** 379–393 (1997).
18. Müller-Wodarg, I. C. F. *et al.* Magnetosphere–atmosphere coupling at Saturn: 1—Response of thermosphere and ionosphere to steady state polar forcing. *Icarus* **221,** 481–494 (2012).
19. Watkins, C. & Cho, J. Y.-K. The vertical structure of Jupiter's equatorial zonal wind above the cloud deck, derived using mesoscale gravity waves. *Geophys. Res. Lett.* **40,** 472–476 (2013).
20. Hickey, M. P., Schubert, G. & Walterscheid, R. L. Acoustic wave heating of the thermosphere. *J. Geophys. Res.* **106,** 21543–21548 (2001).
21. Walterscheid, R. L., Schubert, G. & Brinkman, D. G. Acoustic waves in the upper mesosphere and lower thermosphere generated by deep tropical convection. *J. Geophys. Res.* **108,** 1392 (2003).
22. Schubert, G., Hickey, M. P. & Walterscheid, R. L. Heating of Jupiter's thermosphere by the dissipation of upward propagating acoustic waves. *Icarus* **163,** 398–413 (2003).
23. Flasar, F. M. *et al.* Thermal structure and dynamics of the Jovian atmosphere. I. The Great Red Spot. *J. Geophys. Res.* **86,** 8759–8767 (1981).
24. Fletcher, L. N. *et al.* Thermal structure and composition of Jupiter's Great Red Spot from high-resolution thermal imaging. *Icarus* **208,** 306–328 (2010).

Acknowledgements We thank the Infrared Telescope Facility, which is operated by the University of Hawaii under contract NNH14CK55B with the National Aeronautics and Space Administration (NASA). We are grateful to the observing staff at the Infrared Telescope Facility and Mauna Kea Observatory. This work was funded by NASA under grant number 9500303356 issued through the Planetary Astronomy Program (to L.M. and J.O'D). The UK Science and Technology Facilities Council (STFC) supported this work through the Studentship Enhancement Programme (STEP) for J.O'D., and consolidated grant support for T.S.S. and H.M. (ST/N000749/1). The Royal Astronomical Society partially funded travel to take the observations. We are grateful for the planetary ephemerides that were provided by the Planetary Data System.

Author Contributions J.O'D. collected, analysed and interpreted the data and wrote the paper. L.M. greatly assisted in the data reduction, analysis, interpretation and writing of the paper. T.S.S. helped with the analysis and interpretation of the data. H.M. assisted in the collection and reduction of data, and provided computer code necessary for the analysis of data. All authors provided comments on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.O'D. (jameso@bu.edu).

Reviewer Information *Nature* thanks J. Cho, M. Flasar and J. Sinclair for their contribution to the peer review of this work.

METHODS

Additional observing details. In Fig. 1, where we show the acquisition of Jovian spectra, Jupiter's sub-Earth latitude was +3°. The configuration of the SpeX instrument on the Infrared Telescope Facility was single order with a long slit, at a spectral resolution of $R = 2{,}500$. The slit length and width used were 60 arcsec and 0.3 arcsec, respectively, and one pixel subtended 0.15 arcsec on the sky. In Fig. 2 the model telluric transmission spectrum is obtained from the Atmospheric TRANsmission database (ATRAN; https://atran.sofia.usra.edu) for a spectral resolution of $R = 2{,}500$. The absorption wells near H_3^+ lines in the centre of the spectrum in Fig. 2 serve to highlight our reasons for avoiding that region in the temperature fitting. The attenuation of the signal in this figure by the sky is constant as a function of latitude because all of the temperature fits are from the same exposure, so any attenuation would affect each temperature as a function of latitude in the same way.

Absolute calibration. We flux calibrated the data by using the photometric-standard A0V star HR1019 in the usual manner: that is, by assuming a blackbody curve for the temperature of the star (10,000 K in this case) and comparing it to what we observed. This serves a dual purpose in that by dividing the data by the flux calibration, it converts counts into physical units of flux and also yields a profile of what the sky has absorbed. The mean uncertainty in the absolute calibration as a function of wavelength is 4% of the flux, and the signal-to-noise ratio for the star was 24.

Instrumental effects. These are accounted for by flat fielding, dark-current subtraction and hot pixel removal in every frame. The calibrated Jovian spectra (containing uncertainties in absolute calibration above) also include noise from the instrumentation and Earth's atmospheric attenuation. The uncertainties are thus found by finding the standard deviation of the backgrounds in the final spectrum. All errors are propagated through with the absolute calibration and uncertainty to produce the error bars in intensity displayed in Fig. 2 and the temperature estimates in Fig. 3.

H_3^+ fitting. To find the temperatures from Fig. 1b, we used a spectroscopic H_3^+ line list[25] and the most recent H_3^+ partition function coefficients[26]. The spectrum of H_3^+ can be treated as a sum of Gaussian distribution curves, with each curve a function of temperature. This 'equation of a spectrum' is solved in order to derive the temperature[27]. This technique has been used to derive H_3^+ temperatures on Jupiter, Saturn and Uranus for decades[28], with typical uncertainties of 10%. The fitting routines used are the same as those in previous literature[27], and include a list of over three million ro-vibrational transition lines of H_3^+ (ref. 25). The fitting routine uses the most recent partition function constants to establish a temperature; these constants are applicable for temperatures between 100 K and 10,000 K (whereupon the ion dissociates)[26].

Handling of non-H_3^+ intensity. We now address the possibility of attenuation of H_3^+ by other sources at Jupiter. Possibility 1 is that there is enhanced reflection of sunlight from haze at the location of the GRS, but this is not seen adjacent in wavelength to any lines in Fig. 1 and can consequently be ruled out. Possibility 2 pertains to emission from neutral gases. Only the two intensity peaks overlaid with solid red lines are included in the final fit, though the left peak contained the H_3 lines at 3.38285 μm and 3.38391 μm, whereas the right peak line included 3.45502 μm, 3.45483 μm and 3.45468 μm. Methane (CH_4), the dominant hydrocarbon in Jupiter's atmosphere, is known to emit at a number of wavelengths in this region, namely 3.380 μm, 3.392 μm, 3.404 μm, 3.415 μm, 3.440 μm and 3.454 μm. Some of these are visible in Fig. 1 (for example, 3.404 μm) and some are not (for example, 3.380 μm), but we are mainly interested in any that could affect the fitted H_3^+, which means ignoring, for now, the central portion of Fig. 2. The CH_4 emission line at 3.454 μm is the only line that could possibly fall on a fitted H_3^+ line, and the effect of it doing so would mean that the line ratio between the H_3^+ lines denoted by the solid red fit would be larger. For this particular set of lines, if the ratio is increased, then the temperature estimate decreases: this can be seen by comparing the ratios of lines in Fig. 2, with the lower-ratio GRS spectrum corresponding to 1,644 K ± 161 K, while the higher-ratio non-GRS spectrum is fitted as 900 ± 42 K (standard errors of the mean). In other words, if methane was contributing emission to this line, then accounting for it in some way by removing an arbitrary amount would result in the GRS temperature fitted being even higher than the 1,600 K derived here.

Code availability. The H_3^+ spectroscopic line list used in the model is available online at http://www.exomol.com/data/molecules. In addition, an online H_3^+ intensity calculator is available at http://h3plus.uiuc.edu. The model-fitting routines and reduction code used in this work are available on request from J.O'D. (jameso@bu.edu). Our data reduction pipeline makes substantial use of the NASA Astronomy IDL library, available online at http://idlastro.gsfc.nasa.gov.

25. Neale, L., Miller, S. & Tennyson, J. Spectroscopic properties of the H_3^+ molecule: a new calculated line list. *Astrophys. J.* **464,** 516–520 (1996).
26. Miller, S., Stallard, T., Melin, H. & Tennyson, J. H_3^+ cooling in planetary atmospheres. *Faraday Discuss.* **147,** 283–291 (2010).
27. Melin, H. *et al.* On the anticorrelation between H_3^+ temperature and density in giant planet ionospheres. *Mon. Not. R. Astron. Soc.* **438,** 1611–1617 (2014).
28. Stallard, T. S. *et al.* Temperature changes and energy inputs in giant planet atmospheres: what we are learning from H_3^+. *Phil. Trans. R. Soc.* **370,** 5213–5224 (2012).

LETTER

doi:10.1038/nature18592

A photon–photon quantum gate based on a single atom in an optical resonator

Bastian Hacker[1]*, Stephan Welte[1]*, Gerhard Rempe[1] & Stephan Ritter[1]

That two photons pass each other undisturbed in free space is ideal for the faithful transmission of information, but prohibits an interaction between the photons. Such an interaction is, however, required for a plethora of applications in optical quantum information processing[1]. The long-standing challenge here is to realize a deterministic photon–photon gate, that is, a mutually controlled logic operation on the quantum states of the photons. This requires an interaction so strong that each of the two photons can shift the other's phase by π radians. For polarization qubits, this amounts to the conditional flipping of one photon's polarization to an orthogonal state. So far, only probabilistic gates[2] based on linear optics and photon detectors have been realized[3], because "no known or foreseen material has an optical nonlinearity strong enough to implement this conditional phase shift"[4]. Meanwhile, tremendous progress in the development of quantum-nonlinear systems has opened up new possibilities for single-photon experiments[5]. Platforms range from Rydberg blockade in atomic ensembles[6] to single-atom cavity quantum electrodynamics[7]. Applications such as single-photon switches[8] and transistors[9,10], two-photon gateways[11], nondestructive photon detectors[12], photon routers[13] and nonlinear phase shifters[14–18] have been demonstrated, but none of them with the ideal information carriers: optical qubits in discriminable modes. Here we use the strong light–matter coupling provided by a single atom in a high-finesse optical resonator to realize the Duan–Kimble protocol[19] of a universal controlled phase flip (π phase shift) photon–photon quantum gate. We achieve an average gate fidelity of (76.2 ± 3.6) per cent and specifically demonstrate the capability of conditional polarization flipping as well as entanglement generation between independent input photons. This photon–photon quantum gate is a universal quantum logic element, and therefore could perform most existing two-photon operations. The demonstrated feasibility of deterministic protocols for the optical processing of quantum information could lead to new applications in which photons are essential, especially long-distance quantum communication and scalable quantum computing.

Perhaps the simplest way to realize a photonic two-qubit gate is to overlap two photons in a nonlinear medium. However, it has been argued that this cannot ensure full mutual information transfer between the qubits for reasons of locality and causality[20,21]. Instead, a viable strategy is to keep the two photons separate, change the nonlinear medium using the first photon, use this change to affect the second photon, and, finally, make the first photon interact with the medium again to ensure gate reciprocity. These three sequential interactions enable full mutual information exchange between the two qubits, as is required for a gate, even though the photons never meet directly.

Our experimental realization of a controlled phase flip (CPF) photon–photon gate builds on the proposal by Duan and Kimble[19]. The medium is a single atom strongly coupled to a cavity and the interactions happen upon reflection of each photon off the atom–cavity system[22]. The proposal of ref. 19 considers three reflections, but here we replace the second reflection of the first photon by a measurement of the atomic state and classical phase feedback on the first photon (analogous to a proposal[23] in which the roles of light and matter are interchanged). In practice, this allows us to achieve better fidelities, higher efficiencies and to use a simpler setup compared to that of the proposed scheme[19].

We employ a single ^{87}Rb atom trapped in a three-dimensional optical lattice[24] at the centre of a one-sided optical high-finesse cavity[12] (Fig. 1). The measured cavity quantum electrodynamics parameters for the relevant transition $|\uparrow\rangle = |F=2, m_F=2\rangle \leftrightarrow |e\rangle = |F=3, m_F=3\rangle$ of the D_2 line are $(g, \kappa, \gamma) = 2\pi(7, 2.5, 3)$ MHz. Here, F and m_F are the quantum numbers describing the total atomic angular momentum and its projection onto the quantization axis, respectively, g denotes the atom–cavity coupling constant, and κ and γ are the decay rates of the cavity field and the atomic dipole, respectively. The atom takes on the role of an ancilla qubit, implemented in the basis $|\downarrow\rangle = |F=1, m_F=1\rangle$ and $|\uparrow\rangle$, with the quantization axis along the cavity axis. Both photonic qubits are individually encoded in the polarization using the notation $|\mathrm{L}\rangle$ and $|\mathrm{R}\rangle$ for a left- and a right-handed photon, respectively. They are consecutively coupled into the cavity beam path via a non-polarizing beam splitter (98.5% transmission), which takes on the role of a polarization-independent circulator. The photons as well as the empty cavity are on resonance with the transition $|\uparrow\rangle \leftrightarrow |e\rangle$ at 780 nm. Only the atom in $|\uparrow\rangle$ and the photon in $|\mathrm{R}\rangle$ are strongly coupled, because the $|\downarrow\rangle \leftrightarrow |e\rangle$ transition is detuned by the ground-state hyperfine splitting of 6.8 GHz, and the left-circularly polarized transition $|\uparrow\rangle \leftrightarrow |F=3, m_F=1\rangle$ is shifted out of resonance by a dynamical Stark shift induced by the laser that traps the atom. The strong light–matter coupling between $|\uparrow\rangle$ and $|\mathrm{R}\rangle$ shifts the phase of a reflected photon by π compared to the cases where the atom occupies $|\downarrow\rangle$ or the photon is in $|\mathrm{L}\rangle$. Thus, each reflection constitutes a bidirectional controlled-Z interaction[22] between the atomic and photonic qubit (red boxes in Fig. 2a).

Figure 2a depicts the experimental implementation of the photon–photon gate as a quantum circuit diagram. In short, the protocol starts with arbitrary photonic input qubits $|p_1\rangle$ and $|p_2\rangle$ and with the atom optically pumped to $|\uparrow\rangle$. After this initialization, two consecutive atomic-qubit rotations combined with controlled-Z atom–photon quantum gates are performed. The purpose of the rotations is to maximize the effect of the subsequent gates. Note that up to this point the first photon has the capability to act via the atom onto the second photon. To implement a back-action of the second photon onto the first one, the protocol ends with a measurement of the atomic qubit and feedback onto the first photon. This measurement has the additional advantage that it removes any possible entanglement of the atom with the photons, as required for an ancillary qubit. A longer and detailed stepwise analysis of the above protocol as well as the characterization of the Raman lasers used for the implementation of the atomic-state rotations can be found in the Methods.

To apply this scheme in practice, the qubits have to be stored and controlled in an appropriately timed sequence, as follows. After the first photon p_1 is reflected, it directly enters a 1.2-km-long delay fibre. The delay time of 6 μs is sufficient to allow for reflection of both photons

[1]Max-Planck-Institut für Quantenoptik, Hans-Kopfermann-Strasse 1, 85748 Garching, Germany.
*These authors contributed equally to this work.

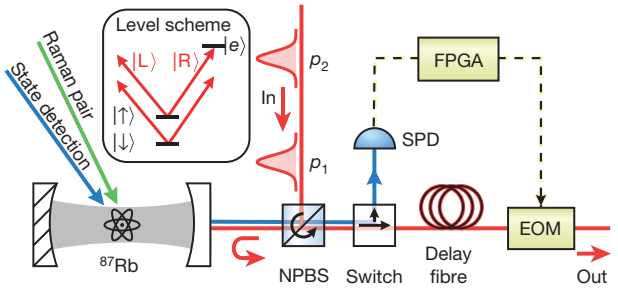


Figure 1 | Schematic of our setup. Qubit-carrying weak coherent photon pulses p_1 and p_2 enter in two separate spatio-temporal modes via a non-polarizing 98.5% transmitting beam splitter (NPBS) that acts effectively as a circulator. The photons are subsequently reflected from the cavity containing a single atom before a switch directs them into a delay fibre. While p_1 and p_2 are stored in the fibre, the state of the atom is read out via fluorescence photons (blue arrows) that the switch directs towards a single-photon detector (SPD). A field programmable gate array (FPGA) applies a conditional phase feedback to p_1 via an electro-optical modulator (EOM). Eventually, the photons leave the gate setup towards polarization analysers. The inset shows the atomic energy level scheme. The three depicted, relevant levels of ^{87}Rb and the photon polarizations are defined in the main text. The photons and the empty cavity are on resonance with the atomic transition $|\uparrow\rangle \leftrightarrow |e\rangle$.

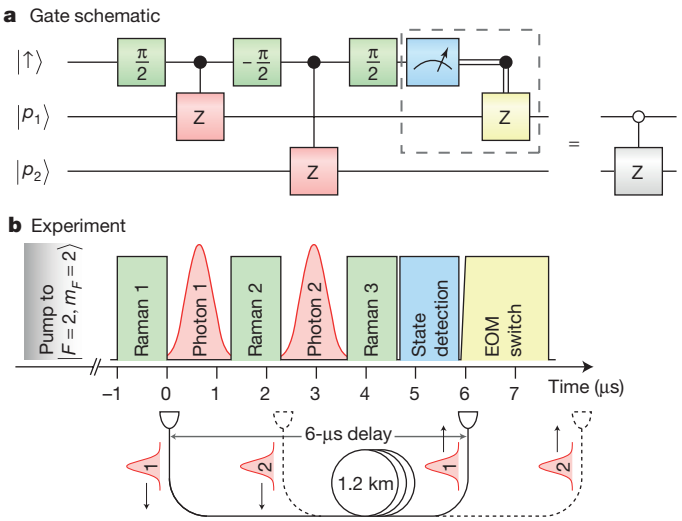


Figure 2 | The photon–photon gate mechanism. a, Quantum circuit diagram. The sequence of controlled-Z gates between the atomic ancilla qubit and the gate photons interleaved with rotations on the atomic qubit acts as a pure CPF gate on the input photon state $|p_1p_2\rangle$. Note that the dashed box is equivalent to the reflection-based quantum controlled-Z gate of the original proposal via the principles of deferred and implicit measurement. **b**, Pulse sequence showing the timing of the experimental steps of the gate protocol. A delay fibre of length 1.2 km is used to store the gate photons for 6 μs.

from the cavity, two coherent spin rotations, and state detection on the atom (Fig. 2b). The two photon wave packets are in independent spatio-temporal modes, which can in principle be arbitrarily shaped. The only requirement is that the frequency spectrum should fall within the acceptance bandwidth of the cavity (0.7 MHz for $\pm 0.1\pi$ phase shift accuracy). We used Gaussian-like envelopes of 0.6 μs full-width at half-maximum (FWHM) within individual time windows of width 1.3 μs, such that the corresponding FWHM bandwidth of 0.7 MHz leads to an acceptable phase-shift spread.

After the last spin rotation, Purcell-enhanced fluorescence state detection of the atomic qubit is performed. This is achieved within 1.2 μs with a laser beam resonant with the $|\uparrow\rangle \leftrightarrow |e\rangle$ transition and impinging perpendicular to the cavity axis (blue beam in Fig. 1). This yields zero fluorescence photons for $|\downarrow\rangle$ and a near-Poissonian-distributed photon number with an average of 4 for $|\uparrow\rangle$, resulting in a discrimination fidelity of 96%. The fluorescence light shares the same spatial mode as the gate photons and needs to be detected before the first photon leaves the delay fibre. Separation of the fluorescence light from the qubit photons is achieved with an efficient free-space acousto-optical deflector (labelled 'Switch' in Fig. 1). Qubit photons pass the deactivated acousto-optical deflector straight towards the delay fibre, whereas state-detection photons are deflected into the first diffraction order directed at a single-photon detector. The corresponding detection events are evaluated in real time by a field programmable gate array, which activates a π phase shift on the $|R\rangle$ component of the first gate photon if the atom was detected in $|\uparrow\rangle$. No phase shift is applied if the atom was found in $|\downarrow\rangle$. This conditional phase shift is performed by an electro-optical modulator with a switching time of 0.1 μs, which is ready when p_1 leaves the delay fibre and is reset before p_2 appears at the end of the fibre. The experiment runs at a rate of 500 Hz, with each execution preceded by atom cooling, atomic state preparation via optical pumping and probing of cavity transmission to confirm success of the initialization. All experiments with one detected qubit photon in each of the two temporal output modes are evaluated without further post-selection.

If both input photons are circularly polarized, the photon–photon gate appears as a CPF gate (see Methods) characterized by:

$$|RR\rangle \rightarrow |RR\rangle \qquad |LR\rangle \rightarrow -|LR\rangle$$
$$|RL\rangle \rightarrow |RL\rangle \qquad |LL\rangle \rightarrow |LL\rangle$$

As with any quantum gate, it can also be expressed in other bases. We define the linear polarization bases as $|H\rangle = \frac{1}{\sqrt{2}}(|R\rangle + |L\rangle)$, $|V\rangle = \frac{1}{\sqrt{2}}(|R\rangle - |L\rangle)$, $|D\rangle = \frac{1}{\sqrt{2}}(|R\rangle + i|L\rangle)$, and $|A\rangle = \frac{-1}{\sqrt{2}}(i|R\rangle + |L\rangle)$, respectively. With one of the photons being circularly and the other one linearly polarized, the gate will act as a controlled-NOT gate with the circular qubit being the control and the linear one being the target qubit. When both photons enter in linear polarization states, the gate will turn the two separable inputs into a maximally entangled state.

We characterized the gate by applying it to various pairs of separable input-qubit combinations and by measuring the average outcome from a large set of repeated trials. The input consisted of two independent weak coherent pulses each impinging with an average photon number of $\bar{n} = 0.17$ onto the cavity. The choice of $\bar{n}$ is a compromise between measurement time and measured gate fidelity. While lowering $\bar{n}$ reduces the data rate because of the high probability of zero-photon events in either of the two photon modes, increasing $\bar{n}$ raises the multi-photon probability per pulse, thereby deteriorating the measured gate fidelity.

First, we processed the four different input states of a controlled-NOT basis, that is, all combinations of photon p_1 in the circular basis and p_2 in a linear basis, and analysed them in the corresponding measurement bases. The resulting truth table is depicted in Fig. 3 and shows an overlap with the case of an ideal controlled-NOT gate of $F_{CNOT} = (76.9 \pm 1.5)\%$.

A decisive property of a quantum gate that distinguishes it from its classical counterpart is its capability to generate entanglement. For both input photons in the linear polarization state $|D\rangle$, the gate ideally creates the maximally entangled Bell state $|\Psi^+\rangle = \frac{1}{\sqrt{2}}(|DL\rangle + |AR\rangle)$. We reconstructed the output of the gate for the input state $|DD\rangle$ from 1,378 detected photon pairs via linear inversion and obtained the density matrix ρ depicted in Fig. 4. It has a fidelity $F_{\Psi^+} = \langle\Psi^+|\rho|\Psi^+\rangle = (72.9 \pm 2.8)\%$ with the ideal Bell state (unbiased linear estimate). The generation of this entangled state from a separable input state directly sets a non-tight bound for the entangling capability (smallest eigenvalue of the partially transposed density matrix)[25] of our gate, $\mathcal{C} \leq -0.242 \pm 0.028$, which is -0.5 for the ideal CPF gate and

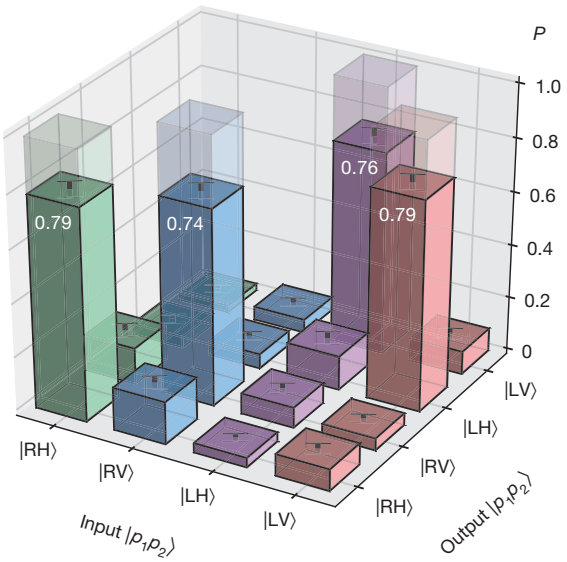


Figure 3 | Truth table of the controlled-NOT photon–photon gate. The gate flips the linear polarization of the target photon p_2 if the control photon p_1 is in the state $|\mathrm{L}\rangle$, but it leaves the target qubit unchanged if the control photon is in $|\mathrm{R}\rangle$. The vertical axis gives the probability of measuring a certain output state given the designated input state. The truth table for an ideal controlled-NOT gate is indicated by the four transparent bars with $P=1$. The black T-shaped bars represent statistical errors (standard error of the mean) on each entry (root mean square 2.2%), computed via linear error propagation assuming independent photon statistics.

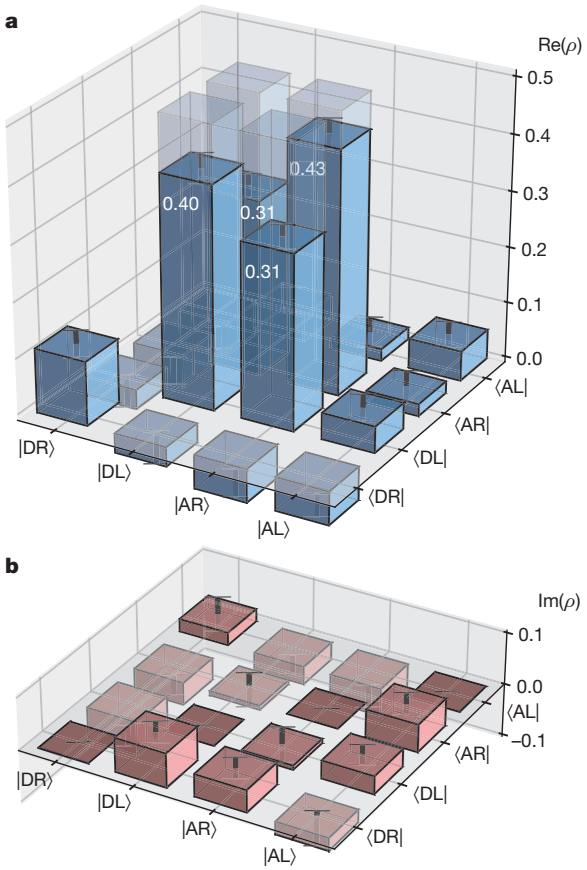


Figure 4 | Reconstructed density matrix of the entangled two-photon state created by the gate from the separable input state $|\mathrm{DD}\rangle$. **a**, **b**, Depicted are the real (**a**) and imaginary (**b**) parts of the elements of the density matrix. The transparent bars indicate the ideal density matrix for $|\Psi^+\rangle$ in the chosen basis. Statistical errors (standard error of the mean) on each entry (root mean square 2.4%) are drawn as black T-shaped bars.

where a negative $\mathcal{C}$ value denotes that the gate is entangling. We remark that the total data set can be separated into two subsets of equal size corresponding to the outcome of the atomic state detection being $|\downarrow\rangle$ or $|\uparrow\rangle$. The respective fidelities are $F^{\downarrow}_{\Psi^+}=(74.4\pm3.9)\%$ and $F^{\uparrow}_{\Psi^+}=(71.5\pm4.2)\%$, that is, the gate works comparably well in both cases.

As an overall measure of the gate performance we determined the average gate fidelity $\overline{F}$, which is equal to the average fidelity of 6×6 output states generated from the input states on all canonical polarization axes (H, V, D, A, R, L) with the theoretically expected ideal outcomes[26]. All 36 state fidelities were estimated linearly and bias-free with randomized tomographically complete basis settings. Although we collected only insignificant statistics of 80 detected photon pairs on each of the output states, their combination gives a meaningful measure of $\overline{F}=(76.2\pm3.6)\%$. The deviation from unity is well understood for our system and results from technical imperfections, which we discuss below.

The efficiency of the presented gate, which is the combined transmission probability for two photons, is unity for the ideal scheme, but gets reduced by several experimental imperfections. It is polarization-independent because all optical elements, including the cavity, have near-equal losses for all polarizations. The two main loss channels are the long delay fibre (transmission $T=40.4\%$) and the limited cavity reflectivity ($R=67\%$). The latter results from the cavity not being perfectly single-sided and having a finite cooperativity of $C=3.3$. All other optical elements have a combined transmission of 81%, dominated by the fibre-coupling efficiency and absorption of the acousto-optical deflector switch. This yields a total experimental gate efficiency of $(22\%)^2=4.8\%$. Despite the transmission losses, characteristic for all photonic devices, the protocol itself is deterministic. The largest potential improvement is offered by eliminating the fibre-induced losses, for instance by a free-space delay line, a delay cavity or an efficient optical quantum memory.

We have modelled all known sources of error (see Methods) to reproduce the deviation of the experimental gate fidelity from unity. Here we quote the reductions in fidelity that each individual effect would introduce to an otherwise perfect gate. The largest contribution stems from using weak coherent pulses to characterize the gate and is therefore not intrinsic to the performance of the gate itself. First, there is a considerable probability of having two photons in one qubit mode if it is populated, resulting in a phase flip of 2π instead of π, causing an overall reduction of the gate fidelity by 12%. Second, the probability of having both qubit modes populated is small, such that detector dark counts contribute a 2% error. The measured gate fidelity could therefore be greatly improved by employing a true single-photon source[7].

The relatively short delay introduced by the optical fibre restricts the temporal windows for the photon pulses and atomic state detection. The resulting bandwidth of the photons reduces the gate fidelity by 6%. The obvious solution is to choose a longer delay. Further errors can be attributed to the characteristics of the optical cavity (5%), the state of the atom (6%), and other optical elements (2%). The cavity has a polarization-eigenmode splitting of 420 kHz that could be eliminated by mirror selection[27]. Neither the resonance frequency of the cavity nor the spatial overlap between its mode and the fibre mode are perfectly controlled (see Methods). The latter could be improved with additional or better optical elements. Fidelity reductions associated with the state of the atom are due to imperfect state preparation, manipulation and detection, and decoherence. Improvements are expected from the application of cavity-enhanced state detection to herald successful state preparation, Raman sideband cooling to eliminate variations in the Stark shift of the atom, and composite pulses to optimize the state rotations. The limited precision of polarization settings and polarization drifts inside the delay fibre are the main contribution from other optical elements. The latter could be improved using active stabilization.

The wealth of realistic suggestions for improvement given above shows that progress towards even higher fidelities is certainly feasible for the gate implementation presented here.

The photon–photon gate as demonstrated here follows a deterministic protocol and could therefore be a scalable building block for new photon-processing tasks such as those required by quantum repeaters[28], for the generation of photonic cluster states[29] or quantum computers[30]. The gate's ability to entangle independent photons could be a resource for quantum communication. Moreover, our gate could serve as the central processing unit of an all-optical quantum computer, envisioned to process pairs of photonic qubits that are individually stored in and retrieved from a quantum cache that may in principle be arbitrarily large. Such a cache would consist of an addressable array of quantum memories, individually connected to the gate via optical fibres. Eventually, such architecture might even be implemented with photonic waveguides on a chip.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 February; accepted 12 May 2016.
Published online 6 July 2016.

1. Kok, P. & Lovett, B. W. *Introduction to Optical Quantum Information Processing* (Cambridge Univ. Press, 2010).
2. Knill, E., Laflamme, R. & Milburn, G. J. A scheme for efficient quantum computation with linear optics. *Nature* **409,** 46–52 (2001).
3. O'Brien, J. L., Pryde, G. J., White, A. G., Ralph, T. C. & Branning, D. Demonstration of an all-optical quantum controlled-NOT gate. *Nature* **426,** 264–267 (2003).
4. O'Brien, J. L. Optical quantum computing. *Science* **318,** 1567–1570 (2007).
5. Chang, D. E., Vuletić, V. & Lukin, M. D. Quantum nonlinear optics—photon by photon. *Nat. Photon.* **8,** 685–694 (2014).
6. Gorshkov, A. V., Otterbach, J., Fleischhauer, M., Pohl, T. & Lukin, M. D. Photon-photon interactions via Rydberg blockade. *Phys. Rev. Lett.* **107,** 133602 (2011).
7. Reiserer, A. & Rempe, G. Cavity-based quantum networks with single atoms and optical photons. *Rev. Mod. Phys.* **87,** 1379–1418 (2015).
8. Baur, S., Tiarks, D., Rempe, G. & Dürr, S. Single-photon switch based on Rydberg blockade. *Phys. Rev. Lett.* **112,** 073901 (2014).
9. Tiarks, D., Baur, S., Schneider, K., Dürr, S. & Rempe, G. Single-photon transistor using a Förster resonance. *Phys. Rev. Lett.* **113,** 053602 (2014).
10. Gorniaczyk, H., Tresp, C., Schmidt, J., Fedder, H. & Hofferberth, S. Single-photon transistor mediated by interstate Rydberg interactions. *Phys. Rev. Lett.* **113,** 053601 (2014).
11. Kubanek, A. *et al.* Two-photon gateway in one-atom cavity quantum electrodynamics. *Phys. Rev. Lett.* **101,** 203602 (2008).
12. Reiserer, A., Ritter, S. & Rempe, G. Nondestructive detection of an optical photon. *Science* **342,** 1349–1351 (2013).
13. Shomroni, I. *et al.* All-optical routing of single photons by a one-atom switch controlled by a single photon. *Science* **345,** 903–906 (2014).
14. Turchette, Q. A., Hood, C. J., Lange, W., Mabuchi, H. & Kimble, H. J. Measurement of conditional phase shifts for quantum logic. *Phys. Rev. Lett.* **75,** 4710–4713 (1995).
15. Tiecke, T. G. *et al.* Nanophotonic quantum phase switch with a single atom. *Nature* **508,** 241–244 (2014).
16. Volz, J., Scheucher, M., Junge, C. & Rauschenbeutel, A. Nonlinear π phase shift for single fibre-guided photons interacting with a single resonator-enhanced atom. *Nat. Photon.* **8,** 965–970 (2014).
17. Beck, K. M., Hosseini, M., Duan, Y. & Vuletić, V. Large conditional single-photon cross-phase modulation. Preprint at https://arxiv.org/abs/1512.02166 (2015).
18. Tiarks, D., Schmidt, S., Rempe, G. & Dürr, S. Optical π phase shift created with a single-photon pulse. *Sci. Adv.* **2,** e1600036 (2016).
19. Duan, L.-M. & Kimble, H. J. Scalable photonic quantum computation through cavity-assisted interactions. *Phys. Rev. Lett.* **92,** 127902 (2004).
20. Shapiro, J. H. Single-photon Kerr nonlinearities do not help quantum computation. *Phys. Rev. A* **73,** 062305 (2006).
21. Gea-Banacloche, J. Impossibility of large phase shifts via the giant Kerr effect with single-photon wave packets. *Phys. Rev. A* **81,** 043823 (2010).
22. Reiserer, A., Kalb, N., Rempe, G. & Ritter, S. A quantum gate between a flying optical photon and a single trapped atom. *Nature* **508,** 237–240 (2014).
23. Duan, L.-M., Wang, B. & Kimble, H. J. Robust quantum gates on neutral atoms with cavity-assisted photon scattering. *Phys. Rev. A* **72,** 032333 (2005).
24. Reiserer, A., Nölleke, C., Ritter, S. & Rempe, G. Ground-state cooling of a single atom at the center of an optical cavity. *Phys. Rev. Lett.* **110,** 223003 (2013).
25. Poyatos, J. F., Cirac, J. I. & Zoller, P. Complete characterization of a quantum process: the two-bit quantum gate. *Phys. Rev. Lett.* **78,** 390–393 (1997).
26. Bagan, E., Baig, M. & Muñoz-Tapia, R. Minimal measurements of the gate fidelity of a qudit map. *Phys. Rev. A* **67,** 014303 (2003).
27. Uphoff, M., Brekenfeld, M., Rempe, G. & Ritter, S. Frequency splitting of polarization eigenmodes in microscopic Fabry-Perot cavities. *New J. Phys.* **17,** 013053 (2015).
28. Briegel, H.-J., Dür, W., Cirac, J. I. & Zoller, P. Quantum repeaters: the role of imperfect local operations in quantum communication. *Phys. Rev. Lett.* **81,** 5932–5935 (1998).
29. Raussendorf, R. & Briegel, H. J. A one-way quantum computer. *Phys. Rev. Lett.* **86,** 5188–5191 (2001).
30. Ladd, T. D. *et al.* Quantum computers. *Nature* **464,** 45–53 (2010).

Acknowledgements We thank N. Kalb, A. Neuzner, A. Reiserer and M. Uphoff for discussions and support throughout the experiment. This work was supported by the European Union (Collaborative Project SIQS) and by the Bundesministerium für Bildung und Forschung via IKT 2020 (Q.com-Q) and by the Deutsche Forschungsgemeinschaft via the excellence cluster Nanosystems Initiative Munich (NIM). S.W. was supported by the doctorate programme Exploring Quantum Matter (ExQM).

Author Contributions All authors contributed to the experiment, the analysis of the results and the writing of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.R. (stephan.ritter@mpq.mpg.de).

METHODS

Composition of the photon–photon CPF gate. The action of the quantum circuit diagram depicted in Fig. 2a can be computed in the eight-dimensional Hilbert space spanned by the atomic ancilla qubit and the two photonic qubits. The atomic single-qubit rotations by $\pi/2$ and $-\pi/2$ are described by the operators $\frac{1}{\sqrt{2}}\begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix}$ and $\frac{1}{\sqrt{2}}\begin{pmatrix}1 & 1\\ -1 & 1\end{pmatrix}$, respectively, in the basis $\{|\uparrow\rangle, |\downarrow\rangle\}$. The atom–photon controlled-Z gate is described by $U_{ap} = \mathrm{diag}(-1, 1, 1, 1)$ in the basis $\{|\uparrow\mathrm{R}\rangle, |\uparrow\mathrm{L}\rangle, |\downarrow\mathrm{R}\rangle\ |\downarrow\mathrm{L}\rangle\}$. As indicated in Fig. 2a, the atom is initially prepared in $|\uparrow\rangle$. Any input state of the two photonic qubits, including entangled states, can be written as

$$|p_1 p_2\rangle = c_{\mathrm{RR}}|\mathrm{RR}\rangle + c_{\mathrm{RL}}|\mathrm{RL}\rangle + c_{\mathrm{LR}}|\mathrm{LR}\rangle + c_{\mathrm{LL}}|\mathrm{LL}\rangle$$

defined by the four complex numbers c_{RR}, c_{RL}, c_{LR} and c_{LL}. Henceforth, we will use the compact notation $|rr\rangle := c_{\mathrm{RR}}|\mathrm{RR}\rangle$, $|rl\rangle := c_{\mathrm{RL}}|\mathrm{RL}\rangle$, $|lr\rangle := c_{\mathrm{LR}}|\mathrm{LR}\rangle$, and $|ll\rangle := c_{\mathrm{LL}}|\mathrm{LL}\rangle$. Therefore, any photon–photon gate operation starts in the collective initial state:

$$|\uparrow\rangle(|rr\rangle + |rl\rangle + |lr\rangle + |ll\rangle)$$

The first $\pi/2$ rotation brings the atom into a superposition:

$$\frac{1}{\sqrt{2}}(|\uparrow\rangle + |\downarrow\rangle)(|rr\rangle + |rl\rangle + |lr\rangle + |ll\rangle)$$

followed by a controlled-Z interaction between the atom and the first photon, which flips the sign of all states with the atom in $|\uparrow\rangle$ and the first photon in $|\mathrm{R}\rangle$:

$$\frac{1}{\sqrt{2}}((-|\uparrow\rangle + |\downarrow\rangle)(|rr\rangle + |rl\rangle) + (|\uparrow\rangle + |\downarrow\rangle)(|lr\rangle + |ll\rangle))$$

Subsequent rotation of the atom by $-\pi/2$ creates the state:

$$|\downarrow\rangle(|rr\rangle + |rl\rangle) + |\uparrow\rangle(|lr\rangle + |ll\rangle)$$

Reflection of the second photon flips the sign of all states with the atom in $|\uparrow\rangle$ and the second photon in $|\mathrm{R}\rangle$:

$$|\downarrow\rangle(|rr\rangle + |rl\rangle) + |\uparrow\rangle(-|lr\rangle + |ll\rangle)$$

The final rotation of the atom by $\pi/2$ yields:

$$\frac{1}{\sqrt{2}}((-|\uparrow\rangle + |\downarrow\rangle)(|rr\rangle + |rl\rangle) + (|\uparrow\rangle + |\downarrow\rangle)(-|lr\rangle + |ll\rangle))$$

At this point the state of the atom is measured. There are two equally probable outcomes projecting the two-photon state accordingly:

$$|\uparrow\rangle: \quad -|rr\rangle - |rl\rangle - |lr\rangle + |ll\rangle$$

and

$$|\downarrow\rangle: \quad +|rr\rangle + |rl\rangle - |lr\rangle + |ll\rangle$$

Following detection of the atom in $|\uparrow\rangle$, an additional π phase is imprinted on the $|\mathrm{R}\rangle$-part of the first photon, that is, a sign flip on $|rr\rangle$ and $|rl\rangle$, whereas the photonic state is left unaltered upon detection of $|\downarrow\rangle$. Thereby, the final photonic state becomes

$$|rr\rangle + |rl\rangle - |lr\rangle + |ll\rangle$$

independent of the outcome of the atomic state detection. It differs from the input state by a minus sign on $|lr\rangle$ only. Hence, the total circuit acts as a pure photonic CPF gate:

$$|\mathrm{RR}\rangle \rightarrow |\mathrm{RR}\rangle \qquad |\mathrm{LR}\rangle \rightarrow -|\mathrm{LR}\rangle$$
$$|\mathrm{RL}\rangle \rightarrow |\mathrm{RL}\rangle \qquad |\mathrm{LL}\rangle \rightarrow |\mathrm{LL}\rangle$$

Calibration of atomic single-qubit rotations. To calibrate the relevant experimental parameters, we employ a Ramsey-like sequence of three subsequent rotation pulses. The pulses are exactly timed as in the gate sequence (see Fig. 2), but the two photon pulses interleaved between the Raman pulses are turned off.

Initially, the atom is prepared in $|\uparrow\rangle$. The Raman pair is red-detuned by 131 GHz from the D_1 line of ^{87}Rb. Employing an acousto-optic modulator, we scan one of the Raman lasers over 2.5 MHz while the frequency of the other is fixed. Thus, we effectively scan the two-photon detuning. Extended Data Fig. 1 shows a spectrum depicting the population in $|\uparrow\rangle$ as a function of the two-photon detuning. Ideally, the gate experiments are performed on two-photon resonance. In this case, the second pulse compensates the first and the third one brings the atom into the superposition state $(|\uparrow\rangle + |\downarrow\rangle)/\sqrt{2}$, such that 50% population in $|\uparrow\rangle$ are obtained.

To determine the experimental parameters that guarantee this situation, a theoretical model is fitted to the spectrum. It allows us to simultaneously access several mutually dependent fit parameters that are useful in calibrating the frequency as well as the intensity of our Raman beams. The fit reveals the Rabi frequency for the transition between $|\downarrow\rangle$ and $|\uparrow\rangle$, which we tune to 250 kHz to obtain $\pi/2$ pulses in 1 μs. The two-photon detuning is also extractable from the fit and we find a light shift of 40 kHz that is due to the Raman lasers. To compensate for it, we choose different two-photon detunings when the pulses are on and off, such that two-photon resonance is guaranteed during the entire sequence.

Transverse optical mode matching. Good overlap between the transverse mode profiles of the incoming wave packet and the optical cavity is essential for the performance of the gate. To achieve this, the qubit-carrying photon pulses are taken from a single-mode fibre with its mode matched to the cavity. In a characterization measurement we determined that 92% of probe light emanating from the cavity is coupled into this input fibre. Therefore, 8% of the impinging light may arrive in an orthogonal mode that does not interact with the atom–cavity system. Light in this mode reduces the fidelity of the gate if it is collected at the output. This problem is overcome because the delay fibre also acts as a filter for the transverse mode profile after the cavity. The mode overlap between cavity and delay fibre is 84%, partially suffering from mode distortion by the acousto-optical deflector used for path switching. From an analysis of cavity reflection spectra we can estimate the amount of light that did not interact with the cavity but is still coupled from the input fibre into the delay fibre. It is below 1% of the gate output, such that the resulting reduction of the gate fidelity is also well below 1%.

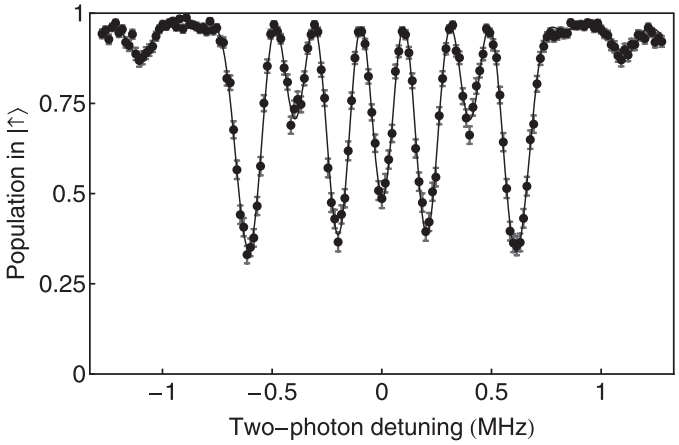
A small misalignment, for example, due to slow temperature drifts, reduces the positive filtering effect described above. Therefore, optimal mode matching is essential to maintain maximum gate fidelity. In the experiment, reflection spectra of the empty cavity were constantly monitored and, whenever necessary, data taking was interrupted to re-establish optimal mode overlap.

Simulation of imperfections. To understand the imperfections encountered in the experiment, we have set up a model of both photonic qubits and the atomic ancilla qubit in terms of their three-particle density matrix ρ. Under ideal conditions, the density matrix transforms via sequential unitary transformations U as $\rho \rightarrow U\rho U^{\dagger}$, and known error sources can be introduced at each specific step. Finally, the fidelity of ρ with the desired target state is calculated for comparison with the experimental value.

In this scenario, an unnoticed, incorrect preparation of the atom creates an incoherent admixture of the wrong initial state. Errors in the atomic state detection lead to an exchange of the photonic submatrices corresponding to each atomic state. Detector dark counts are modelled as an admixture of a fully mixed state and decoherence effects are taken into account as reductions in off-diagonal elements of ρ. Cases where photons do not enter the cavity because of geometric mode mismatch are included with a phase shift of zero, and the case of an undetected additional photon in one of the weak pulses is incorporated with a phase shift of 2π, that is, twice the ideal value. Interestingly, most deteriorations of the atom–photon interaction, like fluctuations of the atomic, cavity and photon frequencies, all condense into a variation, $\Delta\varphi = \pm 0.15\pi$, of the conditional phase shift. Considering this together with the polarization rotation $R_{\mathrm{p}}(\xi)$ that a photon experiences owing to the residual cavity birefringence by an angle of $\xi = 0.06\pi$ in the case of $|\downarrow\rangle$, the ideal atom–photon controlled-Z gate $U_{ap} = \mathrm{diag}(-1, 1, 1, 1)$ in the basis $\{|\uparrow\mathrm{R}\rangle, |\uparrow\mathrm{L}\rangle, |\downarrow\mathrm{R}\rangle, |\downarrow\mathrm{L}\rangle\}$ must be replaced by:

$$U_{\mathrm{ap}} = \begin{bmatrix} e^{i(\pi+\Delta\varphi)} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \multicolumn{2}{c}{\multirow{2}{*}{$R_{\mathrm{p}}(\xi)$}} \\ 0 & 0 & & \end{bmatrix}$$

Random fluctuations in some of the parameters enter our model by integrating the resulting density matrix over the assumed Gaussian distribution function.



Extended Data Figure 1 | Ramsey-like spectrum to calibrate the atomic state rotations. After initialization of the atom in $|\uparrow\rangle$, we perform the same sequence of three Raman pulses as in the gate protocol. The final population in $|\uparrow\rangle$ is determined as a function of the two-photon detuning of the employed Raman pair with respect to the frequency difference between the two atomic qubit states. The solid dots are measured data with statistical error bars (standard error of the mean). The solid line is the fit of a theoretical model based on the sequence of rotations. It yields results for the Rabi frequency of the atomic spin rotation, an offset of the two-photon detuning, as for example, induced by ambient magnetic fields, and the light shift imposed by the Raman laser pair, all with ± 3 kHz precision.

LETTER

doi:10.1038/nature18593

Single-layer MoS_2 nanopores as nanopower generators

Jiandong Feng[1], Michael Graf[1], Ke Liu[1], Dmitry Ovchinnikov[2], Dumitru Dumcenco[2], Mohammad Heiranian[3], Vishal Nandigana[3], Narayana R. Aluru[3], Andras Kis[2] & Aleksandra Radenovic[1]

Making use of the osmotic pressure difference between fresh water and seawater is an attractive, renewable and clean way to generate power and is known as 'blue energy'[1–3]. Another electrokinetic phenomenon, called the streaming potential, occurs when an electrolyte is driven through narrow pores either by a pressure gradient[4] or by an osmotic potential resulting from a salt concentration gradient[5]. For this task, membranes made of two-dimensional materials are expected to be the most efficient, because water transport through a membrane scales inversely with membrane thickness[5–7]. Here we demonstrate the use of single-layer molybdenum disulfide (MoS_2) nanopores as osmotic nanopower generators. We observe a large, osmotically induced current produced from a salt gradient with an estimated power density of up to 10^6 watts per square metre—a current that can be attributed mainly to the atomically thin membrane of MoS_2. Low power requirements for nanoelectronic and optoelectric devices can be provided by a neighbouring nanogenerator that harvests energy from the local environment[8–11]—for example, a piezoelectric zinc oxide nanowire array[8] or single-layer MoS_2 (ref. 12). We use our MoS_2 nanopore generator to power a MoS_2 transistor, thus demonstrating a self-powered nanosystem.

MoS_2 nanopores have already demonstrated better water-transport behaviour than graphene[13,14] owing to the enriched hydrophilic surface sites (provided by the molybdenum) that are produced following either irradiation with transmission electron microscopy (TEM)[15] or electrochemical oxidation[16]. The osmotic power is generated by separating two reservoirs containing potassium chloride (KCl) solutions of different concentrations with a freestanding MoS_2 membrane, into which a single nanopore has been introduced[13]. A chemical potential gradient arises at the interface of these two liquids at a nanopore in a 0.65-nm-thick, single-layer MoS_2 membrane, and drives ions spontaneously across the nanopore, forming an osmotic ion flux towards the equilibrium state (Fig. 1a). The presence of surface charges on the pore screens the passing ions according to their charge polarity, and thus results in a net measurable osmotic current, known as reverse electrodialysis[1]. This cation selectivity can be better understood by analysing the concentration of each ion type (potassium and chloride) as a function of the radial distance from the centre of the pore, as we show here through molecular-dynamics simulations (Fig. 1b).

We fabricated MoS_2 nanopores either by TEM[13] (Fig. 1c) or by the recently demonstrated electrochemical reaction (ECR) technique[16]. With a typical nanopore diameter in the range 2–25 nm, a stable

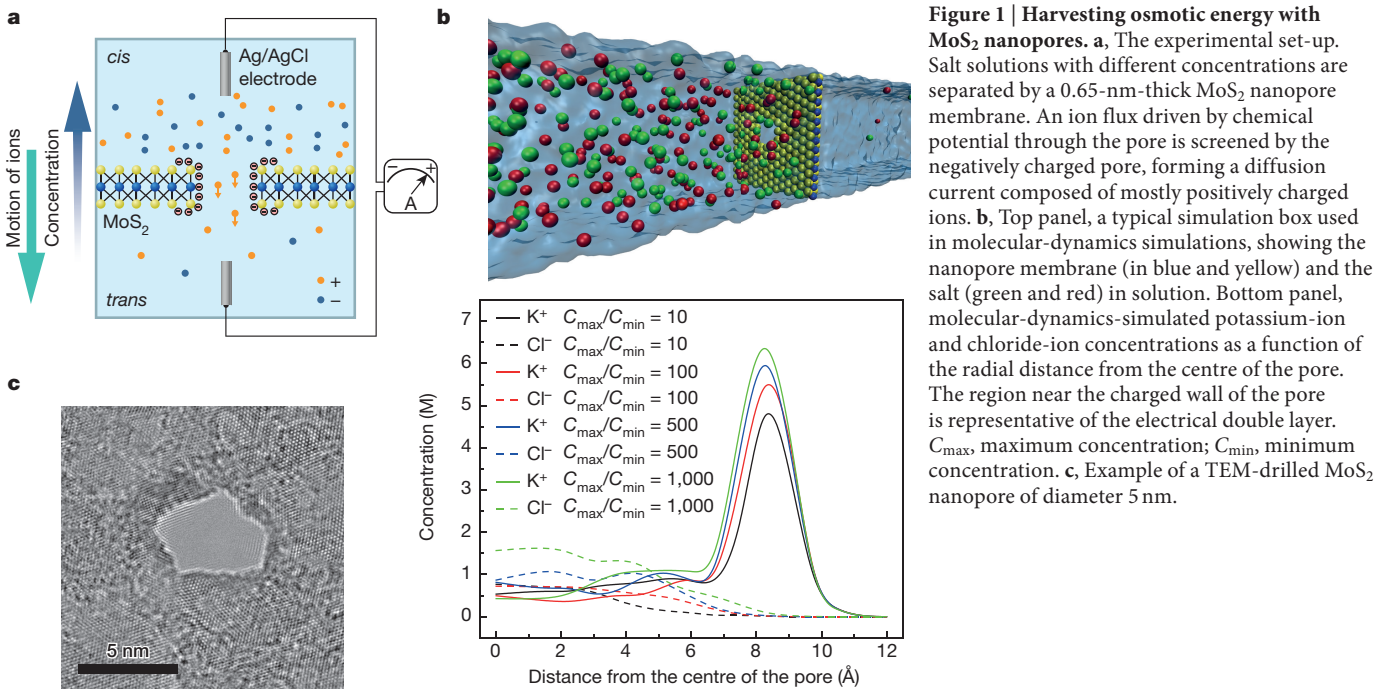


Figure 1 | Harvesting osmotic energy with MoS_2 nanopores. a, The experimental set-up. Salt solutions with different concentrations are separated by a 0.65-nm-thick MoS_2 nanopore membrane. An ion flux driven by chemical potential through the pore is screened by the negatively charged pore, forming a diffusion current composed of mostly positively charged ions. **b**, Top panel, a typical simulation box used in molecular-dynamics simulations, showing the nanopore membrane (in blue and yellow) and the salt (green and red) in solution. Bottom panel, molecular-dynamics-simulated potassium-ion and chloride-ion concentrations as a function of the radial distance from the centre of the pore. The region near the charged wall of the pore is representative of the electrical double layer. C_{max}, maximum concentration; C_{min}, minimum concentration. **c**, Example of a TEM-drilled MoS_2 nanopore of diameter 5 nm.

[1]Laboratory of Nanoscale Biology, Institute of Bioengineering, School of Engineering, EPFL, 1015 Lausanne, Switzerland. [2]Laboratory of Nanoscale Electronics and Structures, Institute of Electrical Engineering and Institute of Materials Science and Engineering, School of Engineering, EPFL, 1015 Lausanne, Switzerland. [3]Department of Mechanical Science and Engineering, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA.

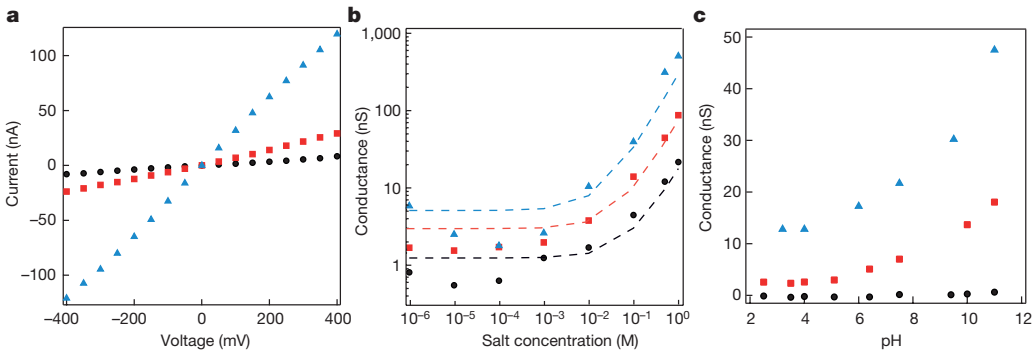


Figure 2 | Electrical conductance and chemical reactivity of the MoS_2 nanopore. a, Current–voltage response of MoS_2 nanopores with different pore sizes (black, 2 nm; red, 6 nm; blue, 25 nm) in 1 M KCl at pH 5. **b**, Conductance as a function of salt concentration at pH 5. By fitting the results to equation (1), we find the extracted surface charge values to be $-0.024\ C\,m^{-2}$, $-0.053\ C\,m^{-2}$ and $-0.088\ C\,m^{-2}$ for a 2-nm, 6-nm and 25-nm pore, respectively. **c**, Conductance as a function of pH for a KCl concentration of 10 mM, for a 2-nm, 6-nm and 25-nm pore.

osmotic current can be expected, owing to the long time required for the system to reach equilibrium. We measured the osmotic current and voltage across the pore by using a pair of Ag/AgCl electrodes to characterize the current–voltage (I–V) response of the nanopore.

To gain a better insight into the performance of the MoS_2 nanopore power generator, we first characterized the ionic transport properties of MoS_2 nanopores under various ionic concentrations and pH conditions, which can provide information on the surface charge of the MoS_2 nanopore. Figure 2a shows the I–V characteristics of MoS_2 nanopores of various diameters. A large pore conductance originates from the ultrathin nature of the membrane. The conductance also depends on the salt concentration (Fig. 2b) and shows saturation at low salt concentrations—a signature of the presence of surface charge on the nanopore[17,18]. The predicted pore conductance (G), taking into account the contribution of the surface charge (Σ), is given by[19]:

$$G = \kappa_b \left[\frac{4L}{\pi d^2} \times \frac{1}{1 + 4\frac{l_{Du}}{d}} + \frac{2}{\alpha d + \beta l_{Du}} \right]^{-1} \quad (1)$$

where κ_b is the bulk conductivity, L is the pore length, d is the pore diameter, l_{Du} is the Dukhin length (which can be approximated by $\frac{|\Sigma|/e}{2c_s}$, where e is the elementary charge and c_s is the salt concentration), α is a geometrical prefactor that depends on the model used (here, $\alpha = 2$)[19], and β can also be approximated to be 2 to obtain the best fitting agreement[19]. From the fitting results shown in Fig. 2b, we find a surface charge value of $-0.024\ C\,m^{-2}$, $-0.053\ C\,m^{-2}$ and $-0.088\ C\,m^{-2}$ for pores of size 2 nm, 6 nm and 25 nm, respectively, at pH 5. These values are comparable to those reported recently for graphene nanopores ($-0.039\ C\,m^{-2}$)[20] and nanotubes ($-0.025\ C\,m^{-2}$ to $-0.125\ C\,m^{-2}$)[5] at pH 5. The surface charge density can be further modulated by adjusting the pH to change the pore surface chemistry (Fig. 2c). The conductance increases with an increase in pH, suggesting the accumulation of more negative surface charges in MoS_2 nanopores. The simulated conductance from equation (1) at 10 mM KCl is linearly proportional to the surface charge values; thus, pH changes could substantially improve the surface charge up to 0.3–0.8 $C\,m^{-2}$. The chemical reactivity of MoS_2 to pH is also supported by measurements of zeta potential on MoS_2 (ref. 21). However we also notice that, as with other nanofluidic systems[5,20], the surface charge density varies from pore to pore, which means that different pores can have disparate values of the equilibrium constant, owing to the various combinations of Mo and S atoms[14] at the edge of the pores (as illuminated by molecular-dynamics simulations[7]).

Next, we introduced a chemical potential gradient by using the KCl concentration gradient system[5]. The concentration gradient ratio is defined as C_{cis}/C_{trans}, where C_{cis} is the KCl concentration in the *cis* chamber and C_{trans} is that in the *trans* chamber; the concentration ranges from 1 mM to 1 M. The highly negatively charged surface selectively passes the ions (in this case potassium ions) according to their polarity, resulting in a net positive current. By measuring the I–V response of the pore in the concentration gradient system (Fig. 3a), we can measure the short-circuit (I_{sc}) current corresponding to zero external bias, while the osmotic potential can be obtained from the open-circuit voltage (V_{oc}). The pure osmotic potential, V_{os}, and current, I_{os}, can then be obtained by subtracting the contribution from the electrode–solution interface at different concentrations; this contribution follows the Nernst equation[5,22] (Extended Data Fig. 1). The osmotic potential is proportional to the concentration gradient ratio (Fig. 3b) and shares a similar trend with the osmotic current (Fig. 3c).

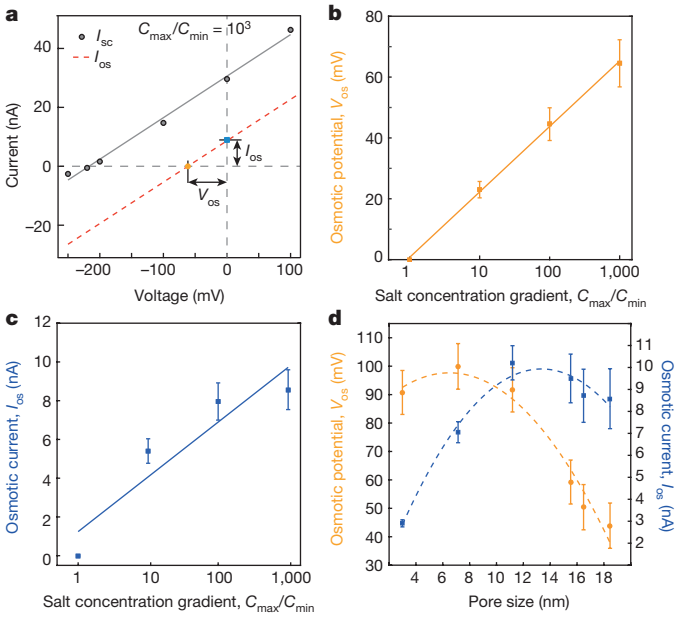


Figure 3 | Osmotic power generation. a, Current–voltage characteristics for a 15-nm nanopore in a 1 M/1 mM KCl gradient. The contribution from the redox reaction on the electrodes is subtracted from the measured total current (grey line) (Extended Data Fig. 1), producing the red dashed line, which represents the pure osmotic contribution. I_{sc} and V_{oc} are the short-circuit current and open-circuit voltage, respectively; I_{os} and V_{os} are the osmotic current and potential. **b**, The generated osmotic potential as a function of the salt gradient. C_{cis} is set to be 1 M KCl; C_{trans} is tunable from 1 mM to 1 M KCl. The solid line represents a linear fitting to equation (2). **c**, Osmotic current as a function of salt gradient. The solid line fits proportionally to the linear part of equation (2). **d**, Osmotic potential and current as a function of pore size. The dashed lines are a guide to the eye and show the trend as the pore size is changed. The error bars come from the corresponding error estimations and represent the s.e.m. (Methods).

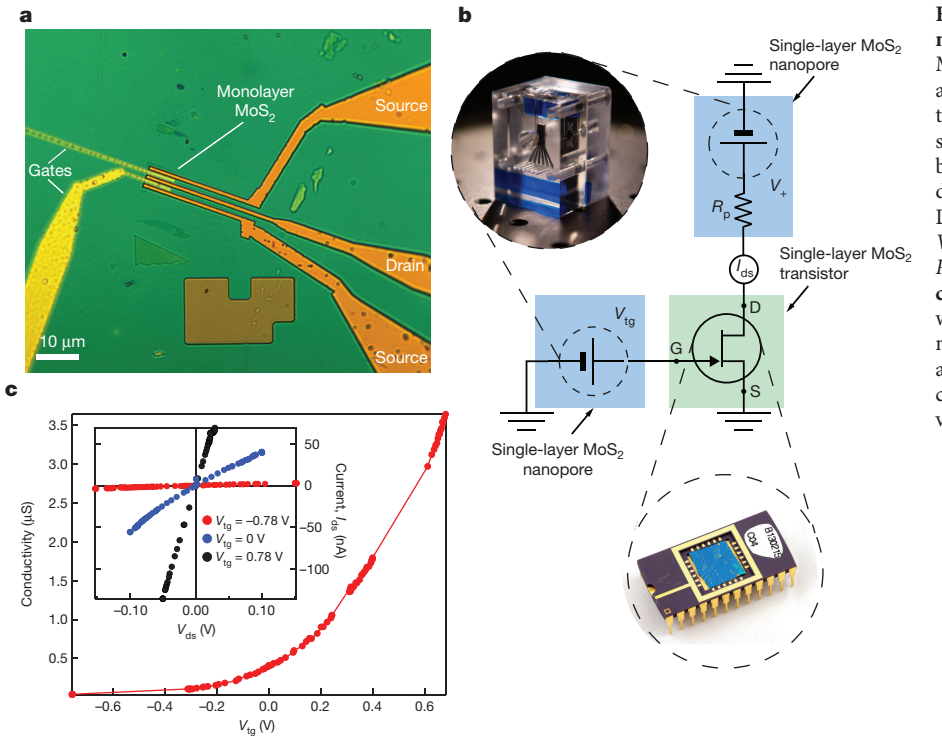


Figure 4 | Demonstration of a self-powered nanosystem. a, Optical image of the fabricated MoS_2 transistor, with a designed gate, and drain and source electrodes. **b**, Circuit diagram for the self-powered nanosystem: the drain–source supply for the MoS_2 transistor is provided by a MoS_2 nanopore, while a second nanopore device operates as the gate voltage source. D, drain; G, gate; S, source; R_p, pore resistance; V_{tg}, gate voltage; V_+, nanopore output voltage. R_p connected in series with V_{tg} has been omitted. **c**, Powering all the terminals of the transistor with nanopore generators. The graph shows the modulated conductivity of the MoS_2 transistor as a function of the top gate voltage (V_{tg}). Inset, current–voltage characteristics at various gate voltages (−0.78 V, 0 V and 0.78 V).

The measured osmotic energy conversion is also pH dependent (Extended Data Fig. 2a, b). The increase in pH leads to higher generated voltage and current, suggesting the importance of surface charge to the ion-selective process.

The extracted osmotic potential is the diffusion potential and it arises from differences in the diffusive fluxes of positive and negative ions, because the pore is ion selective: cations diffuse more rapidly than anions (Fig. 1). The diffusion potential, V_{diff}, can be described as[22]:

$$V_{diff} = S(\Sigma)_{is}\frac{RT}{F}\ln\left[\frac{a_{KCl}^{cis}}{a_{KCl}^{trans}}\right] \quad (2)$$

Here, $S(\Sigma)_{is}$ is the ion selectivity[23] for the MoS_2 nanopore (and equals 1 for the ideal cation-selective case, and 0 for the non-selective case); it is defined as $S(\Sigma)_{is} = t_+ - t_-$, where t_+ and t_- are the transference numbers for positively and negatively charged ions respectively. F, R and T are the Faraday constant, the universal gas constant, and the temperature, respectively, while a_{KCl}^{cis} and a_{KCl}^{trans} are the activities of potassium ions in *cis* and *trans* solutions. By fitting the experimental data presented in Fig. 3b to equation (2), we find the ion-selectivity coefficient $S(\Sigma)_{is}$ to be 0.4, suggesting efficient cation selectivity. This is because the size of our nanopores lies in the range in which the electrical double-layer overlap can occur inside the pore[18], because the Debye length, λ_B, is 10 nm for 1 mM KCl. As shown in Extended Data Fig. 3d, with a concentration gradient of 10 mM/1 mM in a 5-nm pore, the ion selectivity approaches nearly 1, presenting the conditions for ideal cation selectivity[23].

To test the cation-selective behaviour of the pore further, we investigated the relationship between power generation and pore size. As shown in Fig. 3d, small pores display better voltage behaviour, reflecting better performance in terms of ion selectivity. The ion selectivity, $S(\Sigma)_{is}$, decreases from 0.62 to 0.23 as the pore size increases. We calculated the distribution of surface potential for different pore sizes (2 nm, 5 nm and 25 nm) in order to compare the selectivity difference (Extended Data Fig. 3a–c). It has been proven that the net diffusion current stems only from the charge separation and concentration distribution within the electrical double layer[24], and therefore the total current can be expected to increase more rapidly within small pores in the double-layer overlap range compared with larger pore sizes (Fig. 3d). This slight decrease in current in larger pores might be attributed to a reduced local concentration gradient, and also to probable overestimation of the redox potential subtraction. The current can be calculated using either a continuum-based Poisson–Nernst–Planck (PNP) model or molecular-dynamics simulations. The measured dependence of the osmotic potential and osmotic current as a function of the concentration ratio (Fig. 3b, c) is well captured by both computational methods (molecular dynamics, Extended Data Fig. 4, and continuum analysis, Extended Data Fig. 5a). The non-monotonic response to pore size (Fig. 3d and Extended Data Fig. 2c, d) might not only be explained by a possible depletion of the local concentration gradient in large pores, but is also predicted by the continuum-based PNP model (Extended Data Fig. 5b) because of the decrease in ion selectivity.

In order to gain further insight into the thickness scaling, we first verified the pore-conductance relation proposed in equation (1) by using molecular dynamics (Extended Data Fig. 6). We found that ion mobility also scales inversely with membrane thickness (Extended Data Fig. 7a, b), which may conform to previous observations[25]. We then performed molecular-dynamics simulations of multilayer membranes of MoS_2 to investigate the power generated by those membranes. We observe a strong decay in the generated power as the number of layers increases (Extended Data Fig. 7c, d), indicating that the best osmotic power generation occurs in two-dimensional membranes. The consistency between experiments and theoretical models highlights two important factors in achieving efficient power generation from a single-layer MoS_2 nanopore: atomic-scale pore thickness and surface charge.

If we have a single-layer MoS_2 membrane with a homogeneous pore size of 10 nm and a porosity of 30%, then, by exploiting parallelization, the estimated power density would reach 10^6 W m^{-2} with a KCl salt gradient. These values exceed—by two to three orders of magnitude—the results obtained with boron nitride nanotubes[5], and are a million times higher than the power density obtained by reverse electrodialysis with classical exchange membranes[1] (Extended Data Table 1).

As well as using KCl concentration gradients, the nanopore power generator concept could also be applied to liquid–liquid junction

systems with a chemical potential gradient, because the diffusion voltage originates from the Gibbs mixing energy of the two liquids (Supplementary Information). Thus, high-performance, nanopore-based generators based on a large number of available liquid combinations could be explored[24]. For example, we have shown substantial power generation based on a chemical potential gradient that uses two types of liquid (Extended Data Fig. 8d). Considerable energy could also be generated by exploiting parallelization, with multiple small pores or even a continuous porous structure within a large area of single-layer MoS_2 membrane[26], which could be scaled up for mass production using the ECR pore-fabrication technique[16] or plasma-based defect creation[27].

The use of individual nanopores as a micro/nano power source has long been expected[22]. We find here that an individual osmotic generator can also serve as a nanopower source for a self-powered nanosystem, owing to its high efficiency and power density. For this self-powered nanosystem, we chose the high-performance single-layer MoS_2 transistor (Fig. 4a) because of its excellent operation at low power[28]. We characterized this transistor in the configuration shown in Fig. 4b, using two nanopores to apply voltages to the transistor's drain and gate terminals. As shown in Fig. 4c, by varying the top gate voltage in the relatively narrow window of ± 0.78 V, we could modulate the channel conductivity by a factor of 50 to 80. Furthermore, when we fixed the gate voltage and varied the drain–source voltage V_{ds}, (Fig. 4c inset), we obtained a linear I_{ds}–V_{ds} curve, demonstrating efficient injection of electrons into the transistor channel. Further calibration with a standard power source can be found in Extended Data Fig. 8. This system is an ideal self-powered nanosystem in which all the devices are based on single-layer MoS_2.

We have shown that MoS_2 nanopores are promising candidates for investigating osmotic power generation as a renewable energy source. The substantial power generated in our experiments can be attributed mainly to the atomic-scale thickness of the MoS_2 membrane. Our results also provide new avenues for studying other membrane-based processes, such as water desalination[7] or proton transport[29]. Furthermore, the nanopore generator may see application in other ultralow-power devices, such as in electronics.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 6 December 2015; accepted 13 May 2016.
Published online 13 July 2016.

1. Logan, B. E. & Elimelech, M. Membrane-based processes for sustainable power generation using water. *Nature* **488,** 313–319 (2012).
2. Pattle, R. Production of electric power by mixing fresh and salt water in the hydroelectric pile. *Nature* **174,** 660 (1954).
3. Loeb, S. Osmotic power-plants. *Science* **189,** 654–655 (1975).
4. van der Heyden, F. H., Stein, D. & Dekker, C. Streaming currents in a single nanofluidic channel. *Phys. Rev. Lett.* **95,** 116104 (2005).
5. Siria, A. *et al.* Giant osmotic energy conversion measured in a single transmembrane boron nitride nanotube. *Nature* **494,** 455–458 (2013).
6. Suk, M. E. & Aluru, N. Water transport through ultrathin graphene. *J. Phys. Chem. Lett.* **1,** 1590–1594 (2010).
7. Heiranian, M., Farimani, A. B. & Aluru, N. R. Water desalination with a single-layer MoS_2 nanopore. *Nature Commun.* **6,** 8616 (2015).
8. Wang, Z. L. & Song, J. Piezoelectric nanogenerators based on zinc oxide nanowire arrays. *Science* **312,** 242–246 (2006).
9. Wang, Z. L. Self-powered nanosensors and nanosystems. *Adv. Mater.* **24,** 280–285 (2012).
10. Tian, B. *et al.* Coaxial silicon nanowires as solar cells and nanoelectronic power sources. *Nature* **449,** 885–889 (2007).
11. Xu, S. *et al.* Self-powered nanowire devices. *Nature Nanotechnol.* **5,** 366–373 (2010).
12. Wu, W. *et al.* Piezoelectricity of single-atomic-layer MoS_2 for energy conversion and piezotronics. *Nature* **514,** 470–474 (2014).
13. Liu, K., Feng, J., Kis, A. & Radenovic, A. Atomically thin molybdenum disulfide nanopores with high sensitivity for DNA translocation. *ACS Nano* **8,** 2504–2511 (2014).
14. Farimani, A. B., Min, K. & Aluru, N. R. DNA base detection using a single-layer MoS_2. *ACS Nano* **8,** 7914–7922 (2014).
15. Liu, X. *et al.* Top–down fabrication of sub-nanometre semiconducting nanoribbons derived from molybdenum disulfide sheets. *Nature Commun.* **4,** 1776 (2013).
16. Feng, J. *et al.* Electrochemical reaction in single layer MoS_2: nanopores opened atom by atom. *Nano Lett.* **15,** 3431 (2015).
17. Stein, D., Kruithof, M. & Dekker, C. Surface-charge-governed ion transport in nanofluidic channels. *Phys. Rev. Lett.* **93,** 035901 (2004).
18. Bocquet, L. & Charlaix, E. Nanofluidics, from bulk to interfaces. *Chem. Soc. Rev.* **39,** 1073–1095 (2010).
19. Lee, C. *et al.* Large apparent electric size of solid-state nanopores due to spatially extended surface conduction. *Nano Lett.* **12,** 4037–4044 (2012).
20. Shan, Y. *et al.* Surface modification of graphene nanopores for protein translocation. *Nanotechnology* **24,** 495102 (2013).
21. Ge, P. *et al.* Hydrogen evolution across nano-schottky junctions at carbon supported MoS_2 catalysts in biphasic liquid systems. *Chem. Commun.* **48,** 6484–6486 (2012).
22. Kim, D.-K., Duan, C., Chen, Y.-F. & Majumdar, A. Power generation from concentration gradient by reverse electrodialysis in ion-selective nanochannels. *Microfluid. Nanofluidics* **9,** 1215–1224 (2010).
23. Vlassiouk, I., Smirnov, S. & Siwy, Z. Ionic selectivity of single nanochannels. *Nano Lett.* **8,** 1978–1985 (2008).
24. Cao, L. *et al.* Towards understanding the nanofluidic reverse electrodialysis system: well matched charge selectivity and ionic composition. *Energy Environ. Sci.* **4,** 2259–2266 (2011).
25. Wu, J., Gerstandt, K., Zhang, H., Liu, J. & Hinds, B. J. Electrophoretically induced aqueous flow through single-walled carbon nanotube membranes. *Nature Nanotechnol.* **7,** 133–139 (2012).
26. Waduge, P. *et al.* Direct and scalable deposition of atomically thin low-noise MoS_2 membranes on apertures. *ACS Nano* **9,** 7352–7359 (2015).
27. Surwade, S. P. *et al.* Water desalination using nanoporous single-layer graphene. *Nature Nanotechnol.* **10,** 459–464 (2015).
28. Radisavljevic, B., Radenovic, A., Brivio, J., Giacometti, V. & Kis, A. Single-layer MoS_2 transistors. *Nature Nanotechnol.* **6,** 147–150 (2011).
29. Walker, M. I., Braeuninger-Weimer, P., Weatherup, R. S., Hofmann, S. & Keyser, U. F. Measuring the proton selectivity of graphene membranes. *Appl. Phys. Lett.* **107,** 213104 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was financially supported by the European Research Council (grant 259398, PorABEL), by a Swiss National Science Foundation (SNSF) Consolidator grant (BIONIC BSCGI0_157802), by SNSF Sinergia grant 147607, and by funding from the European Union's Seventh Framework Programme FP7/2007-2013 under Grant Agreement 318804 (for single-nanometre lithography). We thank the Centre Interdisciplinaire de Microscopie Electronique (CIME) at the École Polytechnique fédérale de Lausanne (EPFL) for access to electron microscopes. Device fabrication was partially carried out at the EPFL Center for Micro/Nanotechnology (CMi). N.R.A. is supported by the Air Force Office of Scientific Research under grant FA9550-12-1-0464, and by the National Science Foundation under grants 1264282, 1420882, 1506619 and1545907. We acknowledge the use of the parallel computing resource Blue Waters, provided by the University of Illinois and the National Center for Supercomputing Applications.

Author Contributions J.F. and A.R. conceived the idea, designed all experiments, and wrote the manuscript. J.F. and M.G. performed measurements and data analysis. J.F. and K.L. fabricated the nanopore device. D.O. fabricated the MoS_2 transistor and D.D. performed chemical-vapour-deposition MoS_2 growth under A.K.'s supervision. J.F. and D.O. demonstrated the self-powering of the nanosystem. M.H., V.N., and N.R.A. built the computational nanofluidics model and interpreted the simulation results. All authors provided constructive comments on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.F. (jiandong.feng@epfl.ch) or A.R. (aleksandra.radenovic@epfl.ch).

Reviewer Information *Nature* thanks Z. Siwy and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Nanopore fabrication. We fabricated MoS_2 nanopores either by using the atomic-scale ECR technique[16] or by electron irradiation under TEM[13]. Prior to nanopore fabrication, we create a freestanding MoS_2 membrane[30]. Briefly, we use KOH wet etching to prepare SiN_x membranes (of size 10 μm × 10 μm to 50 μm × 50 μm; 20 nm thick). We then use focused ion beam (FIB) or ebeam lithography (followed by reactive ion etching) to drill an opening of 50–300 nm in the membrane. Next we suspend single-layer MoS_2 membranes, grown by chemical-vapour deposition, on the opening by transferring them from sapphire growth substrates[30]. TEM irradiation can be applied to drill a single pore and image the pore. ECR is done by applying a step-like transmembrane potential to the membrane and monitoring the transmembrane ionic current with a Femto DLPCA-200 amplifier (Femto Messtechnik GmbH), with a custom-made feedback control on transmembrane conductance. Nanopores are formed when reaching the critical voltage of MoS_2 oxidation (>0.8 V). We then calibrate the pore size using *I*–*V* characteristics.

Nanofluidic measurements. Nanofluidic transport experiments are performed as described[16]. The nanopore chips are mounted in a custom-made polymethylmethacrylate chamber, and then wetted with an H_2O:ethanol (1:1) solution. Nanofluidic measurements are carried out by taking the *I*–*V* characteristics of the nanopore in different KCl solutions (Sigma Aldrich; the ionic concentration or pH of the solution varies), using an Axopatch 200B patch-clamp amplifier (Molecular Devices Inc.). A pair of chlorinated Ag/AgCl electrodes (which have been rechlorinated regularly) is used to apply voltage and to measure the current. In addition, the electrode potential differences in solutions of different concentrations are calibrated with a saturated Ag/AgCl reference electrode (Sigma Aldrich).

To measure osmotic power generation, we filled the reservoirs with solutions of different concentrations, ranging from 1 mM to 1 M. Measurements are performed at various pH conditions. We found that power generation was optimal at pH 11. First, we measured the *I*–*V* response; we obtained the short-circuit current from the interception of the curve at zero voltage, and the open-circuit voltage from the interception of the curve at zero current. Next, to get the purely osmotically driven contribution, we subtracted the contribution made by the electrode potential difference that results from the redox potential in different concentrations (Extended Data Fig. 1).

For all experiments, we performed cross-checking measurements, including changing the direction of pH and concentration to make sure that the nanopores were not substantially enlarged during the experiments. Most MoS_2 pores were generally stable during hours of experiments owing to their high mechanical strength and stability within the ±600 mV bias range. Thus, we strongly recommend the use of small supporting FIB/ebeam-drilled opening windows (of diameter 50–300 nm) for suspended membranes.

Characterization of single-layer MoS_2 transistors. We fabricated single-layer MoS_2 transistors using a procedure similar to that in ref. 28.

For electrical measurements we used an Agilent 5270B source-meter unit (SMU), an SR-570 low-noise current preamplifier and a Keithley 2000 digital multimeter (DMM; input impedance $>10^{10}\,\Omega$). All measurements were performed in ambient conditions in the dark. An improved efficiency of power conversion in nanopores is obtained by using a combination of pure room-temperature ionic liquids: 1-butyl-2-methylimidazolium hexafluorophosphate (Bmim PF_6) and zinc chloride solution.

We compare the performance of the single-layer MoS_2 transistor in two cases. First, we use two nanopores to apply V_{tg} and V_{ds}, while using a current amplifier and voltmeter to control the current and voltage drop across the device (see Extended Data Fig. 8a). In this case, we use voltage dividers to change the source and gate voltage on the device (not shown in Fig. 4a and Extended Data Fig. 8a). Second, we use the SMU to perform standard two-contact measurements.

Although the characteristics of our transistor are similar in both set-ups, we comment here on the difference detected in the conductivity of the ON state. We attribute it to the slow response of the device in the first case. The change in transistor resistance that occurs when applying gate voltage leads to a change in the impedance of the device and thus a change in the applied effective voltage, V_{dev} (measured with a voltmeter connected in parallel). The nanopore reacts to the change in impedance with a certain stabilization time (from 10 s to 100 s). This appears to be a hysteretic effect and influences the conductivity versus gate-voltage measurements. In the second case, on the other hand, $V_{dev} = V_{ds}$ is constant. There are several secondary effects, which might in turn influence the measured values of two-probe conductivity. In relatively short channel devices, applied V_{ds} might partially contribute to gating of the channel and furthermore to modification of contact resistance. This could be understood by comparing the values of V_{ds} (around 100 mV) and V_{tg} (780 mV). We also do not exclude the possibility of slight doping variations and hysteretic effects that occur because of the filling of trap states inside the transistor channel. However, by driving a device to the ON state and stabilizing the current for a reasonable amount of time, we obtained a very good match in drain–source I_{ds}–V_{ds} characteristics (Extended Data Fig. 8c). We thus conclude that, although there are differences in performance in the two cases, these differences originate mainly from the slow response time of the nanopore.

We extracted the resistance and power of the nanopore by using the ionic liquid Bmim PF_6. By considering the simple resistor network (Extended Data Fig. 8d, inset), we could extract the output power as a function of the load resistance, R_{load}. We fit our dependence according to the following model, which assumes a constant V_{out} and R_{pore}:

$$\text{Power} = \frac{V_{out} R_{load}}{(R_p + R_{load})^2}$$

and found a good fit with $V_{out} = 0.83$ V, which is close to the measured V_{out} of 0.78 V, and with a nanopore impedance, R_p, of 9.4 ± 2.1 MΩ (Extended Data Fig. 8d).

Data analysis. All data analysis has been done using custom-made Matlab (R2016a) code. First, we recorded *I*–*V* characteristics with an Axopatch 200B amplifier, by using either an automatic or a manual voltage switch. We then segmented the current trace into pieces of constant voltage, *V*. We extracted the mean, $\mu(V)$, and standard deviation, $\sigma(V)$, of the stable part of each segment and generated an *I*–*V* plot. The error bars are the standard deviations (see Fig. 3 and Extended Data Fig. 2). All *I*–*V* characteristics were linear. In order to propagate the error correctly, we used a linear fitting method[31]. Using this method, we can extract the a, b, σ_a and σ_b values of the first-order polynomial $I(V) = b\,V + a$. The conductance is the slope, b, of the *I*–*V* curve, and a describes the offset. The height of the error bars reported for conductance measurements is $2\sigma_b$.

We report the osmotic power generation using the osmotic current, I_{os}, and osmotic voltage, V_{os}. Starting from the linear-fit values of the *I*–*V* plot, we can calculate the measured current and voltage: $I_{meas} = a$ and $V_{meas} = a/b$. These measured values have to be adjusted for the electrode potential: $V_{os} = V_{meas} - V_{redox}$ and $I_{os} = (V_{os}/V_{meas}) \times I_{meas}$. Assuming an uncertainty in our estimation of redox potential, σ_{redox}, of 5%, we can propagate the errors using the following formulas[32]:

$$\sigma_{V_{os}} = \sqrt{\left(\frac{1}{b}\sigma_a\right)^2 + \left(\frac{a}{b^2}\sigma_b\right)^2 + \sigma_{redox}^2}$$

$$\sigma_{I_{os}} = \sqrt{\sigma_a^2 + (V_{redox}\sigma_b)^2 + b^2\sigma_{redox}^2}$$

We used these relations to calculate the error bars shown in plots of osmotic voltage and current (Fig. 3 and Extended Data Fig. 2).

Computational simulations. *Molecular-dynamics simulations.* These simulations were performed using the LAMMPS package[33]. A MoS_2 membrane was placed between two KCl solutions as shown in Extended Data Fig. 4a. A fixed graphene wall was placed at the end of each solution reservoir. A nanopore was drilled in MoS_2 by removing the desired atoms. The accessible pore diameter considered in all of the molecular-dynamics simulations is 2.2 nm with a surface charge density of $-0.04694\,C\,m^{-2}$. The system dimensions were 6 nm × 6 nm × 36 nm in the *x*, *y* and *z* directions, respectively. We used the extended simple point charge (SPC/E) water model, and applied the SHAKE algorithm to maintain the rigidity of each water molecule. The Lennard Jones (LJ) parameters are tabulated in Supplementary Table 1. The LJ cut-off distance was 12 Å. The long-range interactions were computed by the particle–particle particle–mesh (PPPM) method[34]. Periodic boundary conditions were applied in the *x* and *y* directions. The system is non-periodic in the *z* direction. For each simulation, first the energy of the system was minimized for 10,000 steps. Next, the system was equilibrated in the isothermic–isobaric (otherwise known as NPT) ensemble for 2 ns at a pressure of 1 atm and a temperature of 300 K to reach the equilibrium density of water. Graphene and MoS_2 atoms were held fixed in space during the simulations. Then, canonical (NVT) simulations were performed, during which the temperature was maintained at 300 K by using the Nosè–Hoover thermostat with a time constant of 0.1 ps (refs 35, 36). Trajectories of atoms were collected every picosecond to obtain the results. For accurate mobility calculations, however, the trajectories were stored every ten femtoseconds.

Continuum model. We also used the continuum–based two-dimensional Poisson–Nernst–Planck (PNP) model. We neglected the contribution of H^+ and OH^- ions in this calculation, as their concentrations are much lower compared with the bulk concentration of the other ionic species (K^+ and Cl^-). Hence, water-dissociation effects are not considered in the numerical model. Further, we assumed that the ions are immobile inside the steric layer and do not contribute to the ionic current. We also did not model the Faradaic reactions occurring near the electrode. Finally, we assumed that the convective component of current originating from the fluid flow is negligible and does not contribute to the non-monotonic osmotic current observed in our experiments. We validated this assumption by performing detailed

all-atom molecular-dynamics simulations and predicted the contribution of electroosmotic velocity in comparison with the drift velocity of the ions.

Under these assumptions, the total flux of each ionic species ($\boldsymbol{\Gamma_i}$) is contributed by a diffusive component resulting from the concentration gradient, and an electrophoretic component arising from the potential gradient, as given by:

$$\boldsymbol{\Gamma_i} = -D_i \boldsymbol{\nabla} c_i - \Omega_i z_i F c_i \boldsymbol{\nabla} \phi$$

where F is Faraday's constant, z_i is the valence of the ith species, D_i is the diffusion coefficient, Ω_i is the ionic mobility, c_i is the concentration and ϕ is the electrical potential. Note that the ionic mobility is related to the diffusion coefficient by Einstein's relation[37], $\Omega_i = \frac{D_i}{RT}$, where R is the ideal gas constant and T is the thermodynamic temperature. The mass transport of each ionic species is:

$$\frac{\mathrm{d}c_i}{\mathrm{d}t} = -\boldsymbol{\nabla} \cdot \boldsymbol{\Gamma_i}$$

The individual ionic current (I_i) across the reservoir and the pore is calculated by integrating their respective fluxes over the cross-sectional area, that is:

$$I_i = \int z_i F \boldsymbol{\Gamma_i} \mathrm{d}S$$

The total ionic current at any axial location is calculated as $\boldsymbol{I} = \sum_{i=1}^{m} z_i F \boldsymbol{\Gamma_i} \mathrm{d}S$, where S is the cross-sectional area corresponding to the axial location and m is the number of ionic species. In order to determine the electric potential along the system, we solve the Poisson equation:

$$\boldsymbol{\nabla} \cdot (\epsilon_r \boldsymbol{\nabla} \phi) = -\frac{\rho_e}{\epsilon_0}$$

where ϵ_0 is the permittivity of free space, ϵ_r is the relative permittivity of the medium and ρ_e is the net space charge density of the ions, defined as:

$$\rho_e = F \sum_{i=1}^{m} z_i c_i$$

We provide the necessary boundary conditions for the closure of the problem. The normal flux of each ion is assumed to be zero on all the walls so that there is no leakage of current. To conserve charge on the walls of the pore, the electrostatic boundary condition is given by:

$$\boldsymbol{n} \cdot \boldsymbol{\nabla} \phi = \frac{\sigma}{\epsilon_0 \epsilon_r}$$

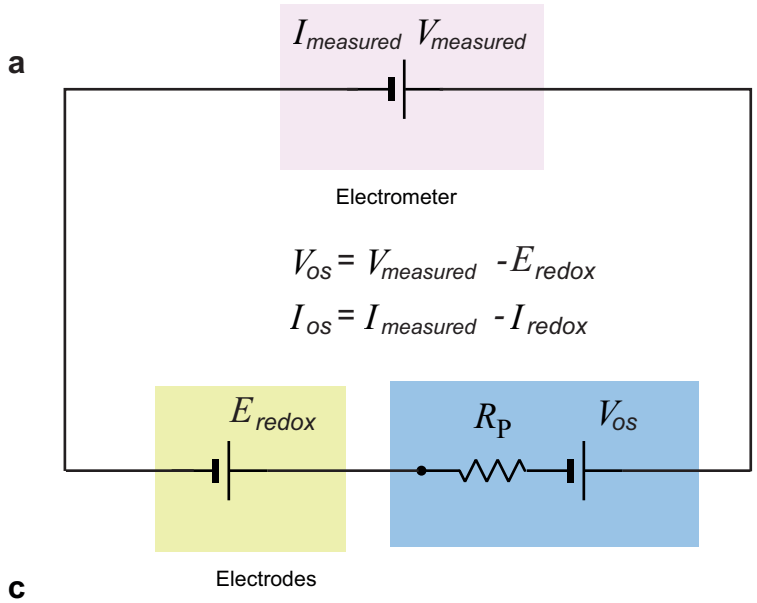
where $\boldsymbol{n}$ denotes the unit normal vector (pointing outwards) to the wall surface and σ is the surface charge density of the walls. The bulk concentration of the *cis* reservoir is maintained at C_{max} and the bulk concentration on the *trans* reservoir is maintained at C_{min}. As we are interested in understanding the osmotic short-circuit current, I_{sc}, we do not apply any voltage difference across the reservoirs. Thus, the boundary conditions at the ends of the *cis* and *trans* reservoirs are specified as:

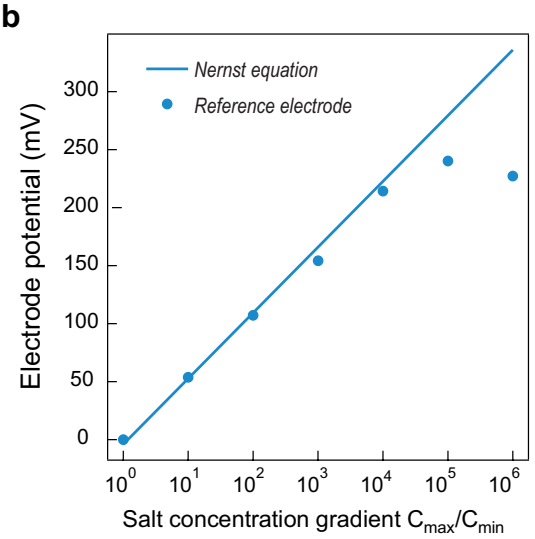
$$c_i = C_{max}, \phi = 0$$

$$c_i = C_{min}, \phi = 0$$

The coupled PNP equations are numerically solved using the finite volume method in OpenFOAM (http://www.openfoam.com/). The details of solver implementation are discussed in refs 38–40. The simulated domain consisted of a MoS_2 nanopore of length, L_n, 0.6 nm and diameter, d_n, varying from 2.2 nm to 25 nm. The simulated length of the reservoir was $L_{cis} = L_{trans} = 11$ nm; the diameter of the reservoir was 50 nm. KCl buffer solution was used in the simulation. The bulk concentration of the *cis* reservoir was fixed at 1 M and the concentration in the *trans* reservoir was varied systematically varied from 1 mM to 1 M. The simulation temperature was 300 K. The bulk diffusivities of K^+ and Cl^- were $1.96 \times 10^{-9}\,m^2\,s^{-1}$ and $2.03 \times 10^{-9}\,m^2\,s^{-1}$. The dielectric constant of the aqueous solution was assumed to be 80. We also assumed zero surface charge density on the walls of the reservoir, as the reservoir is too far away from the nanopore to have an influence on the transport. Unless otherwise stated, the charge on the walls of the MoS_2 nanopore is assumed to be $\sigma_n = -0.04694\,C\,m^{-2}$, consistent with the surface charge calculated from our molecular-dynamics simulations.

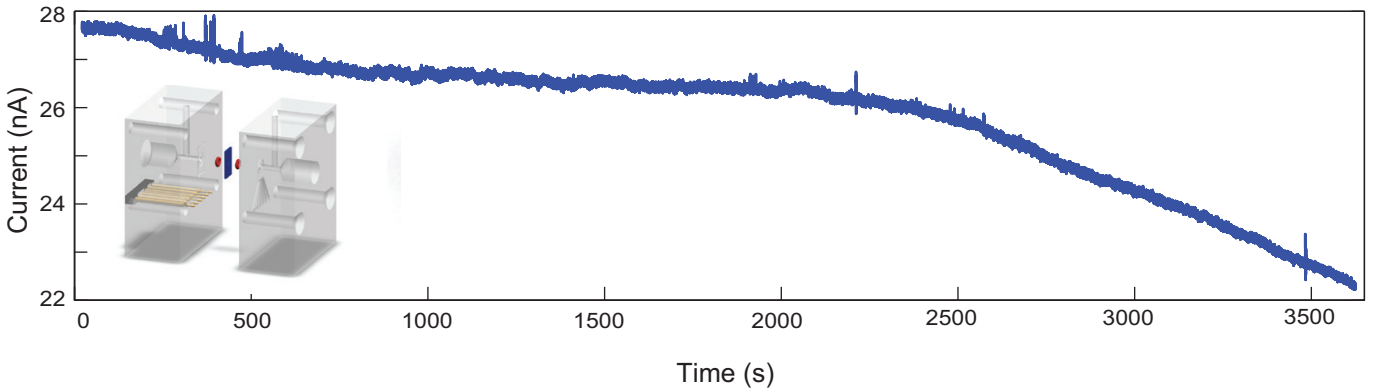
30. Dumcenco, D. *et al.* Large-area epitaxial monolayer MoS_2. *ACS Nano* **9,** 4611–4620 (2015).
31. York, D., Evensen, N. M., Martınez, M. L. & Delgado, J. D. B. Unified equations for the slope, intercept, and standard errors of the best straight line. *Am. J. Phys.* **72,** 367–375 (2004).
32. Ku, H. Notes on the use of propagation of error formulas. *J. Res. National Bureau Standards* **70C,** 263–273 (1966).
33. Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **117,** 1–19 (1995).
34. Hockney, R. W. & Eastwood, J. W. *Computer Simulation Using Particles* (CRC Press, 1988).
35. Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* **81,** 511–519 (1984).
36. Hoover, W. G. Canonical dynamics: equilibrium phase-space distributions. *Phys. Rev. A* **31,** 1695–1697 (1985).
37. Probstein, R. F. *Physicochemical Hydrodynamics: An Introduction.* (John Wiley & Sons, 2005).
38. Nandigana, V. V. & Aluru, N. Understanding anomalous current–voltage characteristics in microchannel–nanochannel interconnect devices. *J. Colloid Interface Sci.* **384,** 162–171 (2012).
39. Nandigana, V. V. & Aluru, N. Nonlinear electrokinetic transport under combined ac and dc fields in micro/nanofluidic interface devices. *J. Fluids Eng.* **135,** 021201 (2013).
40. Nandigana, V. V. & Aluru, N. Characterization of electrochemical properties of a micro–nanochannel integrated system using computational impedance spectroscopy (cis). *Electrochim. Acta* **105,** 514–523 (2013).
41. Weinstein, J. N. & Leitz, F. B. Electric power from differences in salinity: the dialytic battery. *Science* **191,** 557–559 (1976).
42. Audinos, R. Reverse electrodialysis. Study of the electric energy obtained by mixing two solutions of different salinity. *J. Power Sources* **10,** 203–217 (1983).
43. Turek, M. & Bandura, B. Renewable energy by reverse electrodialysis. *Desalination* **205,** 67–74 (2007).
44. Suda, F., Matsuo, T. & Ushioda, D. Transient changes in the power output from the concentration difference cell (dialytic battery) between seawater and river water. *Energy* **32,** 165–173 (2007).
45. Veerman, J., De Jong, R., Saakes, M., Metz, S. & Harmsen, G. Reverse electrodialysis: comparison of six commercial membrane pairs on the thermodynamic efficiency and power density. *J. Membr. Sci.* **343,** 7–15 (2009).



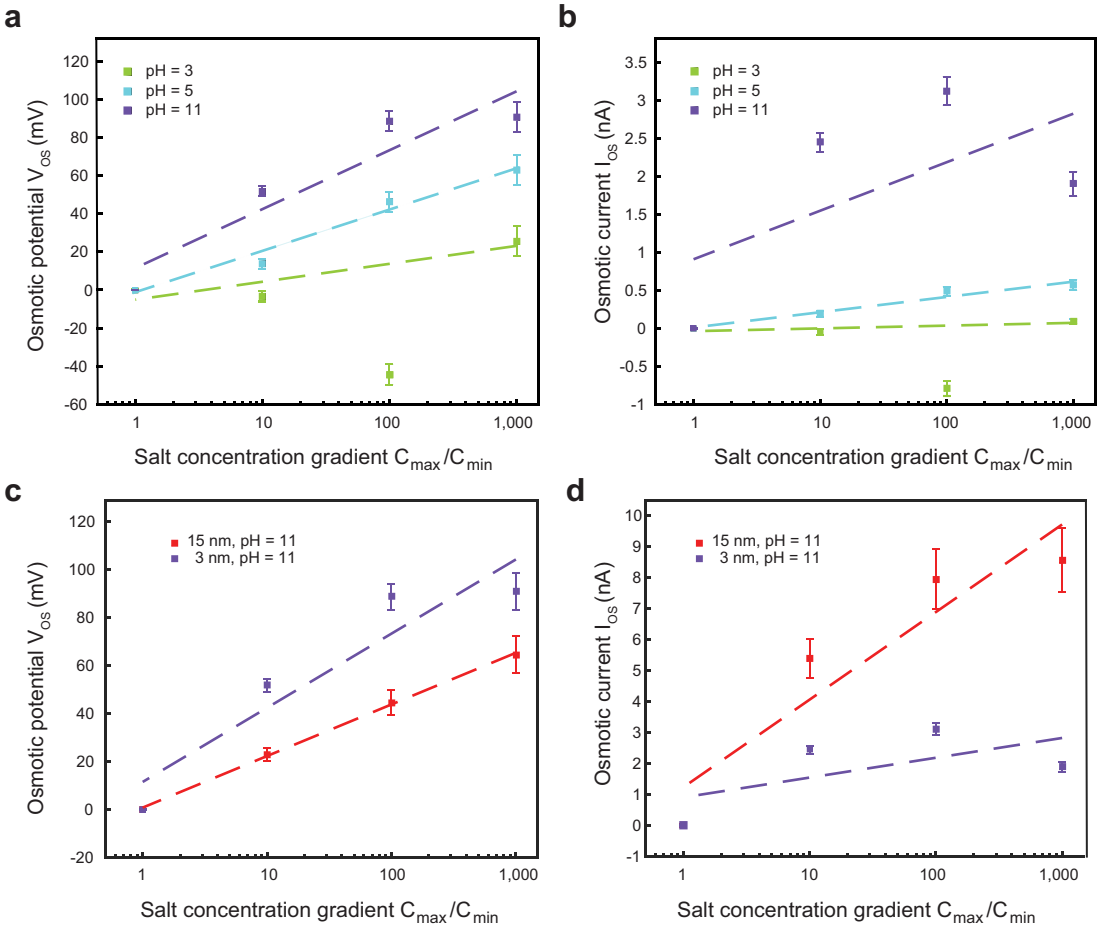


c

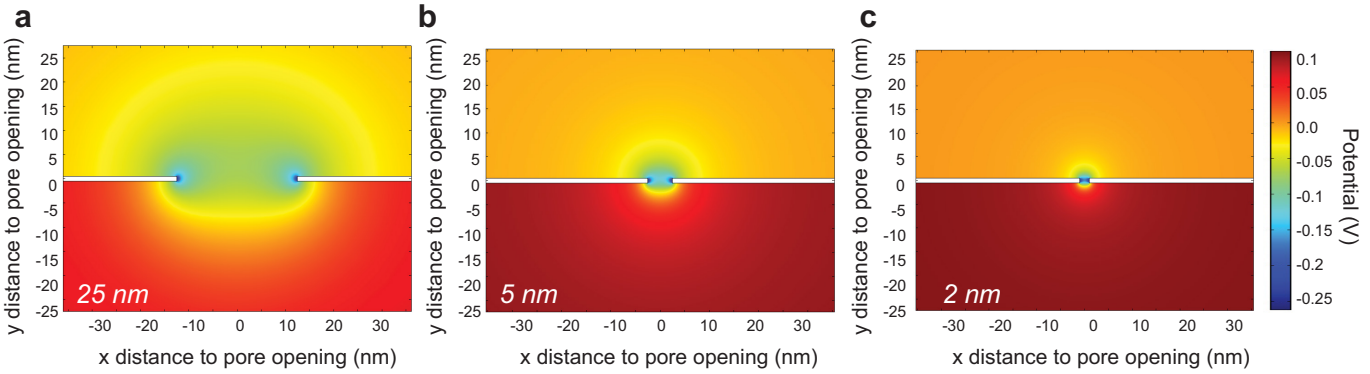
Log C_{max}/C_{min}	$I_{measured}$, nA	$V_{measured}$, mV	E_{redox}, mV	V_{os}, mV	I_{os}, nA
0	Set to 0				
1	18.2	78.2	53.7	24.5	5.7
2	27.2	150.2	107.2	43.0	7.8
3	30.6	216.8	154.1	62.7	8.8
4	20.7	276.0	214.2	61.8	4.6
5	14.5	322.2	240.2	82.0	3.7
6	10.9	311.7	227.0	84.7	3.0



Extended Data Figure 1 | Subtraction of the contribution made by electrodes, and stability of the nanopore generator. a, Diagram showing the contributions of different parts of the system to the overall measured current. The osmotic contribution is obtained by subtracting the contribution of the potential difference at the electrodes from the measured voltage or current. $V_{measured}$ is the measured voltage; E_{redox} is the redox potential difference. **b**, Electrode contribution as a function of the salt concentration gradient: values obtained from the Nernst equation, and measured electrode redox potential differences at the reference electrode. **c**, The data used for the subtraction. E_{redox}, the redox potential at the electrodes. **d**, A 1-hour trace of ionic current, showing the stability of a 14-nm pore in 1 M KCl/1 mM KCl. Inset, the design of the fluidic cell.



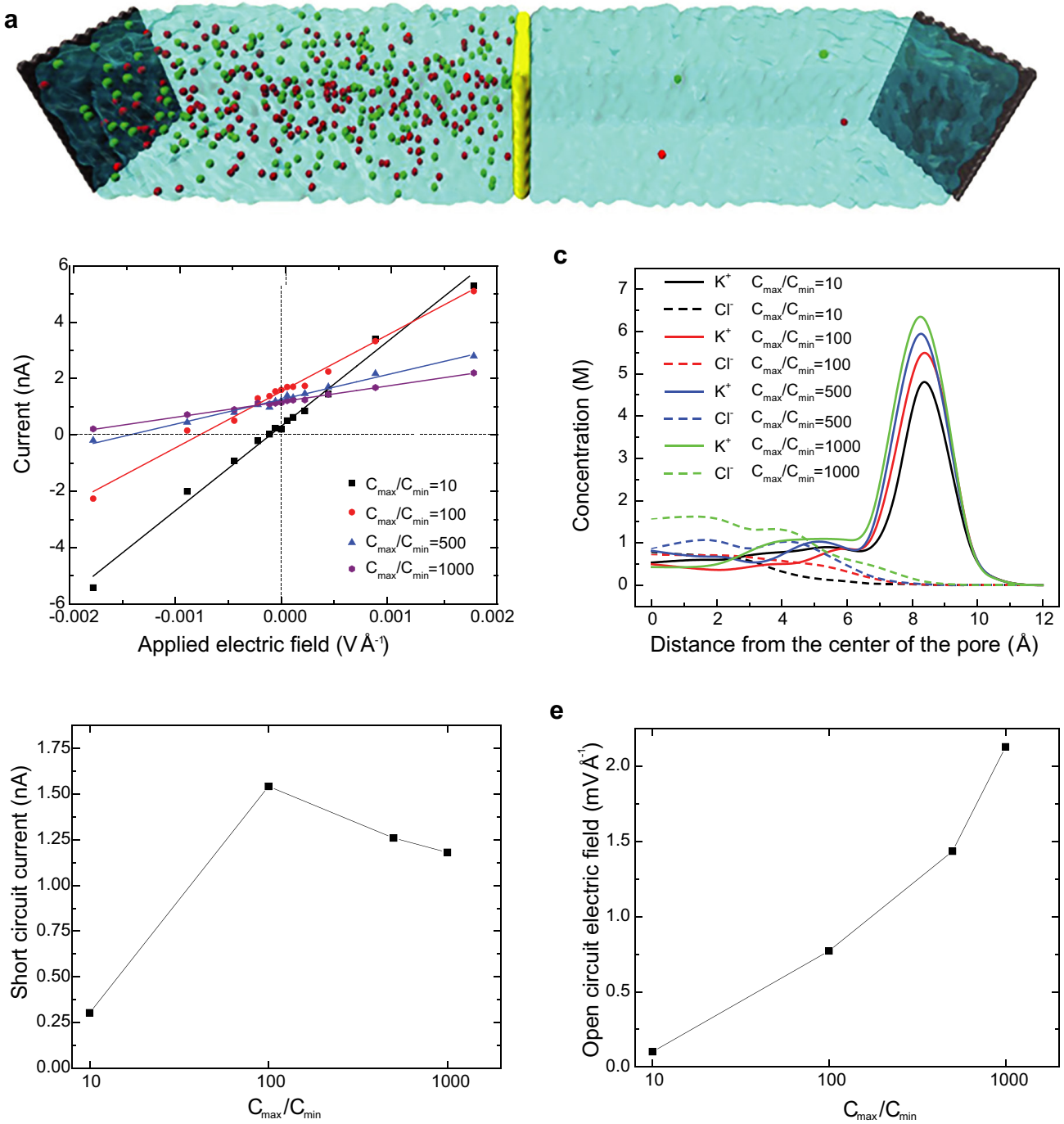
Extended Data Figure 2 | Dependence of osmotic power generation on pH and pore size. a, **b**, Osmotic potential (**a**) and osmotic current (**b**) generated using a 3-nm pore under different pH conditions (pH 3, 5 or 11) and in different concentration gradients. Power generation (both osmotic potential and osmotic current) at pH 3 is very low and sometimes fluctuates to negative, indicating that the pore charge is relatively low. One possible explanation for the negative voltage point is that the surface charge on the pore has fluctuated to positive. **c**, **d**, Osmotic potential (**c**) and osmotic current (**d**) generated using two different pores (3-nm and 15-nm) at pH 11 in different concentration gradients.



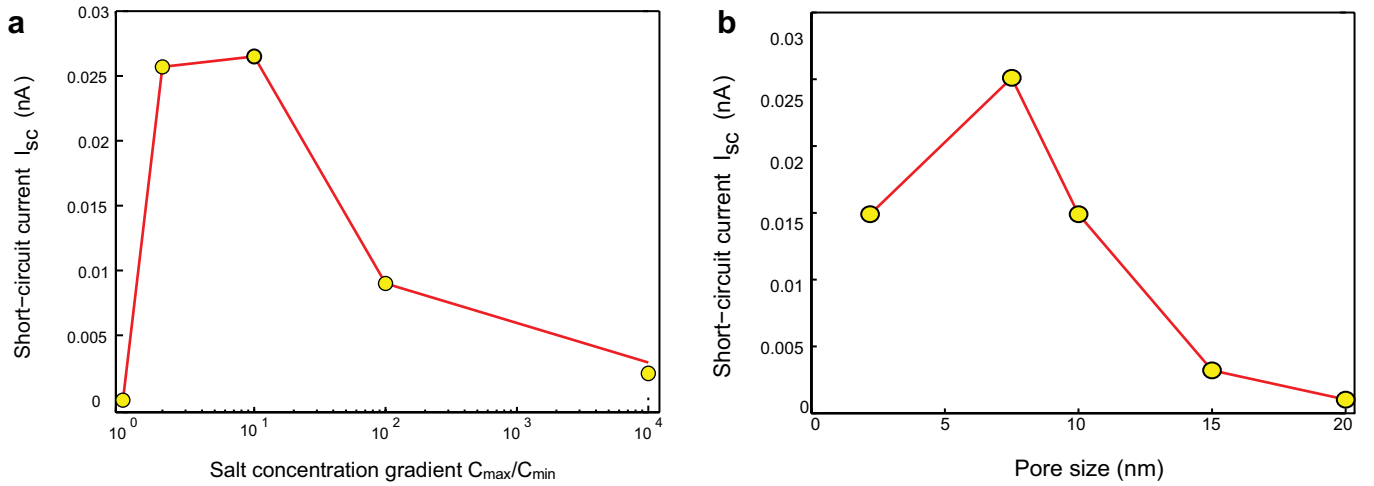
d

C_{min}/C_{max}	$V_{measured}$, mV	E_{redox}, mV	V_{os}, mV	Ion selectivity
1mM/10mM	100.6	46.9	53.7	0.92
10mM/100mM	104.4	53.5	50.9	0.86
10mM/1M	153.3	107.2	46.1	0.78
1mM/100mM	183.0	100.4	82.6	0.7
100mM/1M	67.3	53.7	13.6	0.23

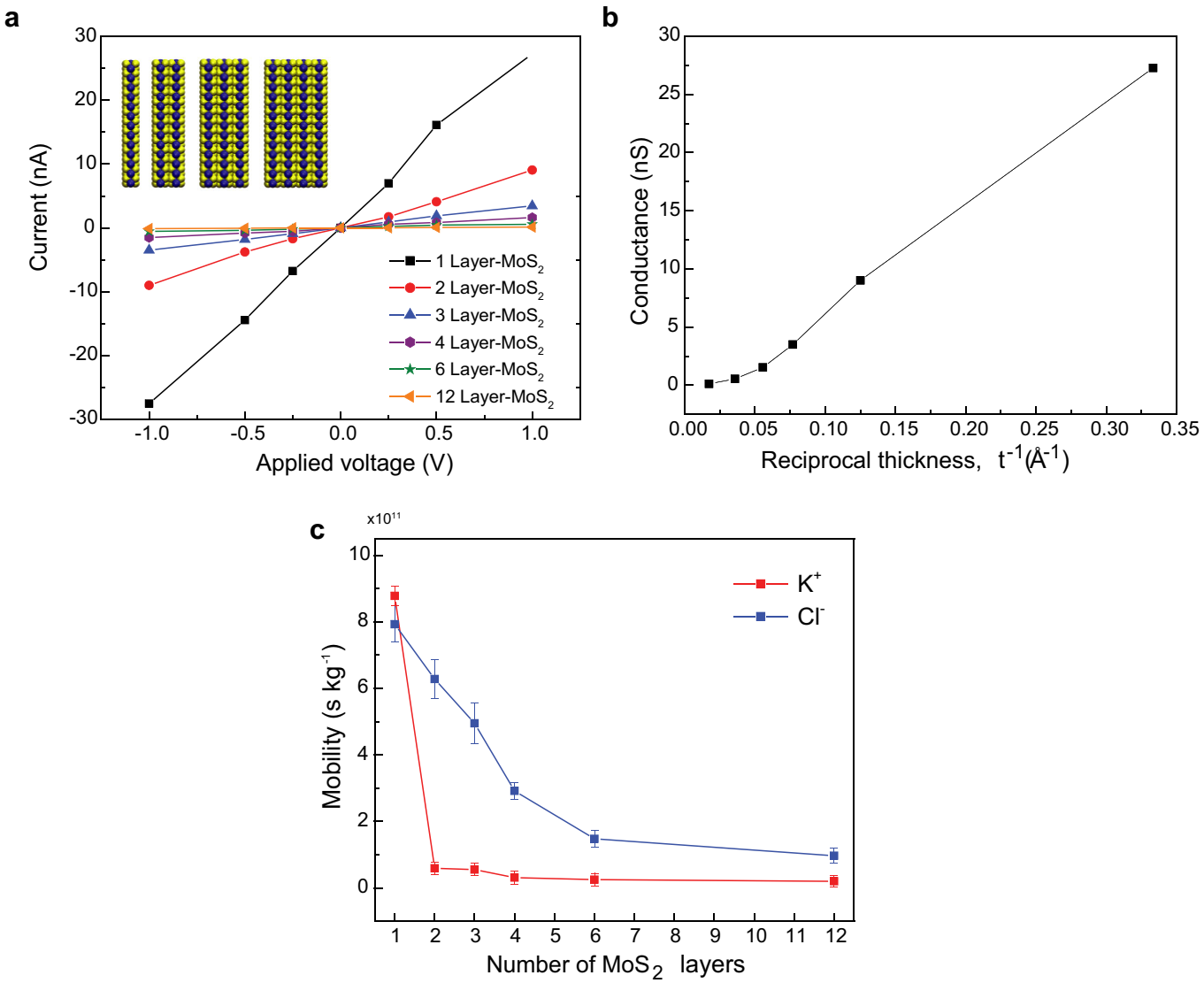
Extended Data Figure 3 | Ideal cation selectivity of the pore. **a**–**c**, Calculated surface potential distribution of MoS_2 nanopores of diameter 25 nm (**a**), 5 nm (**b**), and 2 nm (**c**) under a fixed surface charge density. **d**, Ion selectivity in different salt gradients. The ion selectivity also depends on the Debye length when the concentration gradient is fixed; with a gradient of 10 mM/1 mM and a 5-nm pore, the ion selectivity approaches nearly 1, indicating the ideal cation selectivity.



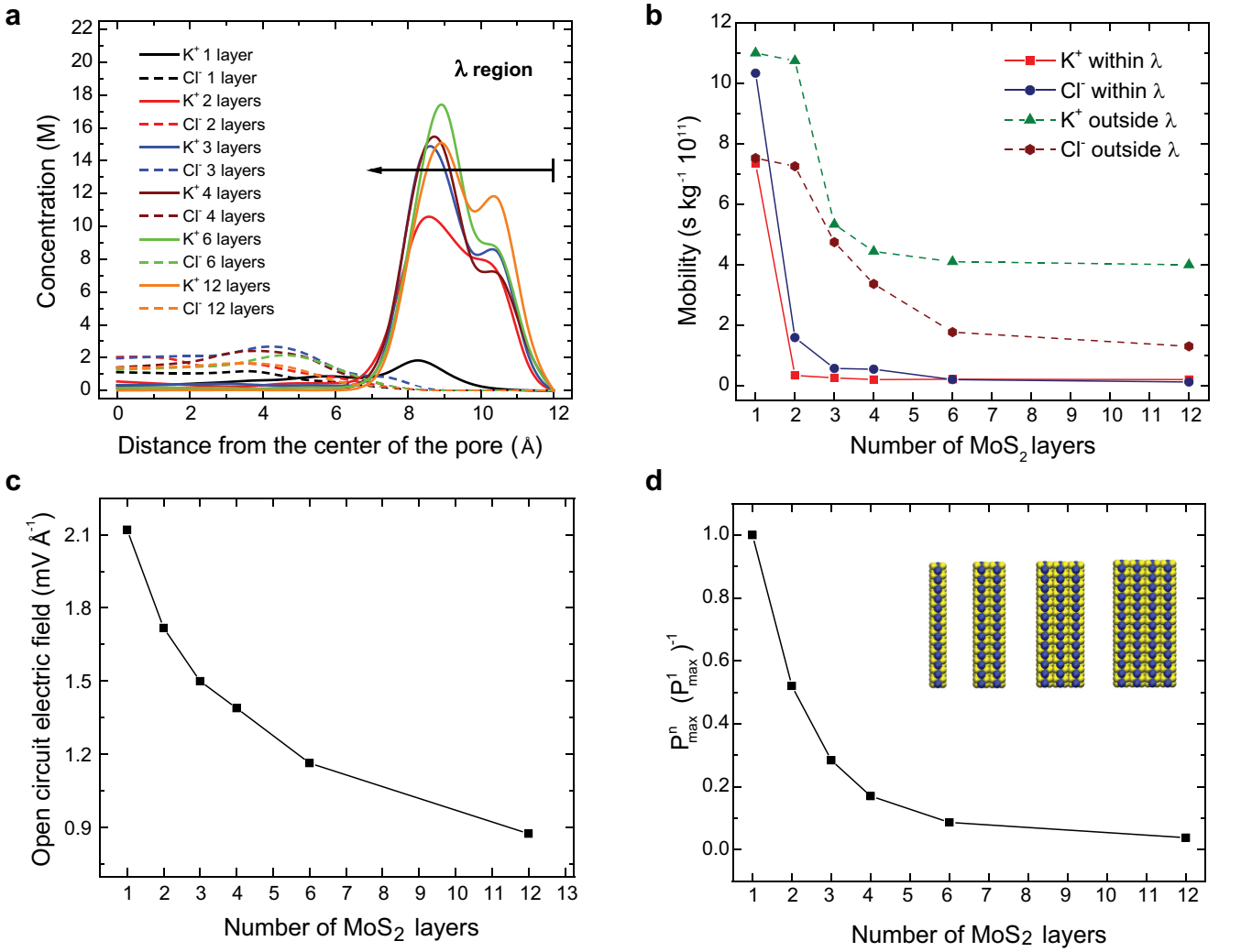
Extended Data Figure 4 | Molecular-dynamics simulations of power generation for various ratios of concentration gradients. a, A typical simulation box. **b**, Current as a function of the applied electric field for a single-layer MoS_2, at different concentration ratios. **c**, K^+ and Cl^- concentrations as a function of the radial distance from the centre of the pore, for different concentration ratios. **d**, Short-circuit current as a function of the concentration ratio. **e**, Open-circuit electric field as a function of the concentration ratio.



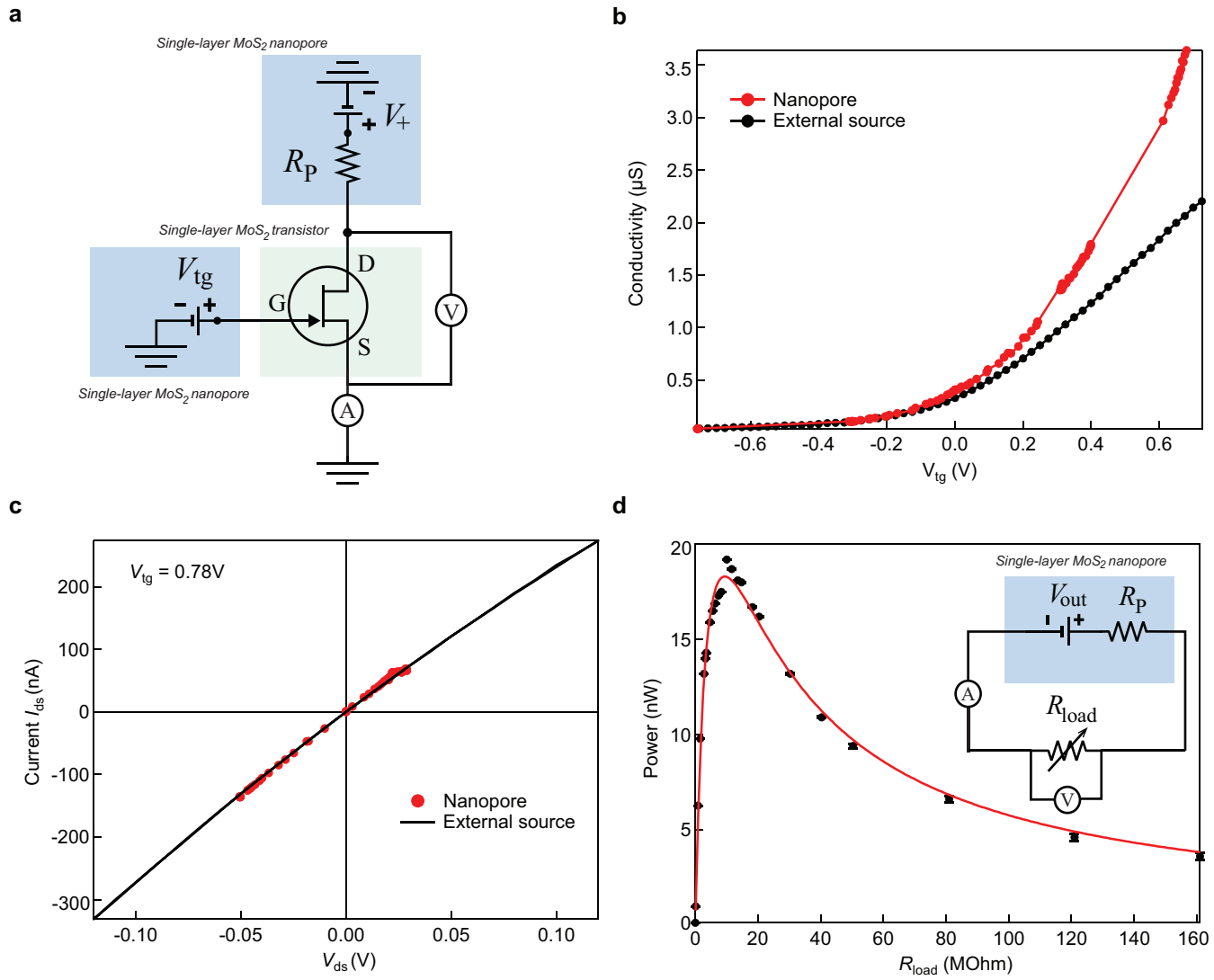
Extended Data Figure 5 | Continuum-based PNP modelling of power generation. a, Short-circuit current, I_{sc}, as a function of the concentration gradient ratio. The diameter of the nanopore here is 2.2 nm. **b**, I_{sc} as a function of the nanopore diameter. The salinity concentration ratio is fixed at 1,000. The surface charge of the nanopore, σ_n, is $-0.04694\ \mathrm{C\,m^{-2}}$.



Extended Data Figure 6 | Molecular-dynamics-modelled conductance as a function of membrane thickness. a, *I*–*V* curves for six membranes with a different number of MoS_2 layers, across a symmetrical 1 M KCl solution. The inset illustrates simulated multilayer membranes. **b**, Conductance of the nanopore as a function of the reciprocal thickness of the membrane (t^{-1}). **c**, Average mobility of each ion for different numbers of layers of MoS_2 membranes.



Extended Data Figure 7 | Simulated power generation as a function of membrane thickness. a, K^+ and Cl^- concentrations as a function of the radial distance from the centre of the pore for single-layer and multilayer membranes. The λ region, near the charged wall of the pore, is representative of the electrical double layer. **b**, The mobility of each ion type within and outside the λ region for different layers of membranes. **c**, The open-circuit electric field across the membrane for different numbers of MoS_2 layers. **d**, Ratio of the maximum power (P) generated by multilayer membranes to the maximum power generated by a single-layer membrane, for different numbers of layers.



Extended Data Figure 8 | Characterization of a single-layer MoS_2 transistor with nanopores and SMU. a, Electrical measurements with two nanopores (V_+, nanopore output voltage; V_{ds}, drain–source voltage; V_{tg}, top gate voltage). The voltage drop across the transistor channel is monitored with the voltmeter (V); current is measured with current amplifier (A). **b**, Comparison of nanopore measurements with standard two-probe measurements made with an external source. **c**, I–V characteristics at $V_{tg} = 0.78$ V after current stabilization, measured in both set-ups. **d**, Output power of nanopore in Bmim PF_6/zinc chloride as a function of load resistance, R_{load}. Inset, circuit diagram for these measurements.

Extended Data Table 1 | Power generation according to membrane thickness

Reverse electrodialysis cells	Power density (W/m^2)	Membrane thickness
Ref. 41	0.17	1 mm
Ref. 42	0.40	3 mm
Ref. 43	0.46	0.19 mm
Ref. 44	0.26	1 mm
Ref. 45	0.95	0.2 mm
Ref. 22	7.7	0.14 mm
Ref. 5	4000	1 μm
This work	10^6	0.65 nm
Multilayer MoS_2 (Simulations)	30000	7.2 nm

The table shows the power generated by membranes of different thickness; data from refs 5, 22, 41–45.

LETTER

doi:10.1038/nature18319

Abrupt plate accelerations shape rifted continental margins

Sascha Brune[1,2], Simon E. Williams[2], Nathaniel P. Butterworth[2] & R. Dietmar Müller[2]

Rifted margins are formed by persistent stretching of continental lithosphere until breakup is achieved. It is well known that strain-rate-dependent processes control rift evolution[1,2], yet quantified extension histories of Earth's major passive margins have become available only recently. Here we investigate rift kinematics globally by applying a new geotectonic analysis technique to revised global plate reconstructions. We find that rifted margins feature an initial, slow rift phase (less than ten millimetres per year, full rate) and that an abrupt increase of plate divergence introduces a fast rift phase. Plate acceleration takes place before continental rupture and considerable margin area is created during each phase. We reproduce the rapid transition from slow to fast extension using analytical and numerical modelling with constant force boundary conditions. The extension models suggest that the two-phase velocity behaviour is caused by a rift-intrinsic strength–velocity feedback, which can be robustly inferred for diverse lithosphere configurations and rheologies. Our results explain differences between proximal and distal margin areas[3] and demonstrate that abrupt plate acceleration during continental rifting is controlled by the nonlinear decay of the resistive rift strength force. This mechanism provides an explanation for several previously unexplained rapid absolute plate motion changes, offering new insights into the balance of plate driving forces through time.

Rifted continental margins, with an overall length of more than 100,000 km, are the longest tectonic features on our planet, two times longer than spreading ridges or convergent plate boundaries. During formation of rifted margins, new continental surface area is generated by normal faulting and volcanic intrusions. Both processes are dependent on extension velocity, which governs the thermal configuration of the rift and hence the depth of the brittle–ductile transition and the length of normal faults[1], as well as the degree of decompression melting and serpentinization[4]. Moreover, rift velocity has been shown to control rift symmetry and the formation of hyper-extended crust[5].

Quantifying the history of extension velocity at rifted margins requires knowledge of the motions between diverging plates and of the timing of continental breakup. Recently, revised regional syn-rift plate models for the opening of the Atlantic, South China Sea, Gulf of California and Australia–Antarctica rifting have become available (see Supplementary Table 1). Here, we incorporate these regional studies in a global plate kinematic model[6] and use an updated, simplified global set of boundaries between continental and oceanic crust (COBs) to explore continental breakup processes. We exploit the fact that in a pre-rift reconstruction, present-day COBs from conjugate passive margins will show substantial overlap, since the plate tectonic models do not explicitly incorporate lithospheric deformation. As the plates move apart the overlap decreases, and the time when conjugate COBs disconnect defines breakup, the transition from rifting to sea-floor spreading (Fig. 1).

We extracted the local rift velocity via pyGPlates, a novel Python library that allows script-based access to the plate reconstruction software GPlates. We subdivided each COB in segments of ~50 km length, and computed the relative velocity between the two contributing plates at each segment (Fig. 1c). First we address the question of whether there is a systematic trend in the temporal evolution of extension velocity within entire rift systems. We visualize the velocity evolution of the rift system in a single diagram by displaying the integrated rift axis length of all segments that deform within a certain velocity range in 1 million year (Myr) time intervals (Fig. 1d). We discarded any segment where breakup is accomplished, that is, where COBs do not overlap anymore. Hence, the analysed plate boundary length declines through time (see dashed grey line in Fig. 1e) and reduces to zero at final continental separation. We explicitly excluded failed rifts from our analysis, because they do not contribute to passive margin formation.

In addition to rift velocity we computed the rate of rifted margin formation, that is, the product of the extension velocity and the velocity-orthogonal length of each COB segment. Integration along both conjugate margins and division by 2 yields the overall rate of rifted margin formation F. The formation rate F increases with extension velocity and decreases when individual segments are discarded during diachronous breakup (Fig. 1e). Note that F is independent of the distance between hinge line and COB, instead representing the newly created margin surface. This allows us to draw robust conclusions on rifted margin growth as no assumptions about previous rift phases, or initial crustal and lithospheric thickness have to be made.

With a rift length of more than 10,000 km, the South Atlantic Rift (Fig. 1) generated 2.1×10^6 km^2 of rifted margin area, more than any other Phanerozoic continental rift. During the first 25 Myr of extension, the Euler pole is located close to the equatorial Atlantic Rift[7,8]; therefore rifting is faster in the southern South Atlantic. The mean rift velocity remains relatively low (<10 mm yr^{-1}, full rate) until it increases rapidly to more than 35 mm yr^{-1} within 6 Myr. This speed-up at ~126–120 Myr ago coincides with severe loss of strength in the equatorial Atlantic Rift[9], suggesting rift weakening as a controlling parameter. Both rift phases shape the rifted margins: about one-third of the South Atlantic margin area was formed during the slow, and two-thirds during the fast phase (Fig. 1e).

This two-phase velocity history is a common feature of many other rift systems, illustrated for the central North Atlantic, North America–Iberia, Australia–Antarctica and South China Sea rifting in Fig. 2 and for the Gulf of California, the northeast Atlantic and North America–Greenland in Extended Data Figs 3 and 4. Consistently, the fast rift phase starts ~10 Myr before inception of breakup and persists until plate separation is complete. Both the slow and the fast phase contribute markedly to shaping the rifted margins. All regional tectonic reconstructions used here, compiled from a range of independent studies, result in the same two-phase pattern, underlining the robustness of our results. Moreover, using alternative reconstructions for the South Atlantic does not change our conclusions (Extended Data Fig. 8).

We conducted robustness tests for all case studies using alternative COBs defined by the extreme landward limit of basement that is not clearly continental crust. These COB polygons are located closer to the coastline, which shortens the duration of rifting by several million

[1]GFZ German Research Centre for Geosciences, Telegrafenberg, 14473 Potsdam, Germany. [2]EarthByte Research Group, School of Geosciences, 2006 University of Sydney, Sydney, Australia.

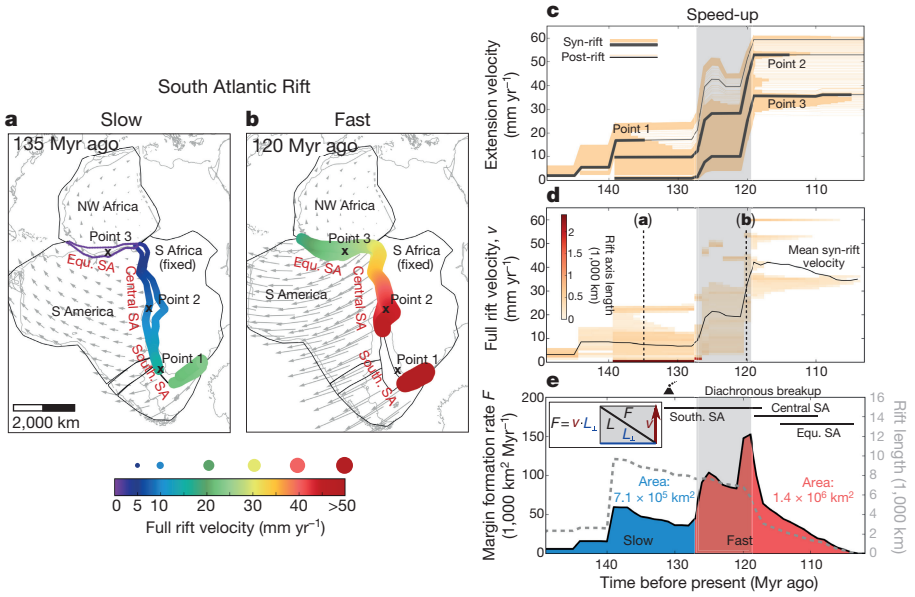


Figure 1 | Rift velocity evolution of the South Atlantic. a, **b**, Rift velocities (coloured circles) are evaluated at overlapping plate polygons (black) with rift-ward polygon boundaries being defined through present-day COBs. Central SA, Central South Atlantic; Equ. SA, equatorial South Atlantic; NW Africa, Northwest Africa; S Africa, South Africa; S America, South America; South. SA, southern South Atlantic. **c**, Extension velocity evolution of all rift points showing the syn-rift and post-rift phase (thick and thin lines, respectively). **d**, Frequency of syn-rift velocities in terms of rift axis length. Colours display the integrated length of all rift segments deforming at the same velocity. The black line shows mean velocity of all syn-rift segments. The fast rift phase starts ~126 Myr ago. **e**, The margin formation rate F at each rift segment of length L is computed by multiplying rift velocity v and velocity-orthogonal length of the segment $L_\perp$. Total rift length is shown as grey dashed line. Timing of Paraná–Etendeka flood basalts is depicted by volcano symbol. Black horizontal bars indicate diachronous breakup. For an animation of the kinematic evolution, see Supplementary Video 1.

years. Nevertheless, the contributions of both phases in shaping the rifted margins remain substantial (Extended Data Figs 5 and 6).

To evaluate the geodynamic response of a rift zone to plate driving forces, we used two-dimensional analytical and numerical models. While the most common modelling approach is to prescribe a constant velocity at the model boundary[5,10], here we use a force boundary condition[11–13] allowing for self-consistent computation of velocity evolution. The force boundary condition is applicable to major rifts where the integrated strength of the entire rift system is comparable to the plate driving forces, such as those considered here.

First, we developed an analytical model of a homogeneous lithospheric layer that deforms according to power law rheology under a constant force. For simplicity, we neglect depth-dependent thinning, non-constant temperature and compressibility; however, these processes are incorporated within the numerical models described later. The resulting velocity is $v = L/(n\,t_r) \times (1 - t/t_r)^{-1}$, where L is the width of the necking zone, n the dislocation creep stress-exponent, t_r the duration of rifting and t is time (see Methods for derivation). The formula is plotted in Fig. 3c using typical parameter values of $L = 100$ km and $t_r = 20$ Myr. Realistic values[14] for n range between 3 and 4, hence we approximate a purely viscous lithosphere by $n = 3.5$ (model A1). Introducing brittle failure as a highly nonlinear end-member of power law creep[12], we also show solutions for $n = 10$ (model A2) and $n = 30$ (model A3). In all cases, rifting commences slowly, yet velocities increase within a few million years, where the short duration of speed-up is caused by the power law rheology of the lithosphere.

Numerical modelling accounts for more realistic evolution of deformation, temperature and buoyancy, but the resulting velocity evolution is very similar to the analytical solution (Fig. 3c). The reason for the acceleration is a feedback mechanism that we term dynamic rift weakening: as long as the rift is strong, the extension rate is low (Fig. 3a), but with continued deformation the rift centre becomes successively weaker due to necking and strain softening (Fig. 3b). Loss of strength accelerates rifting, which results in new strength loss

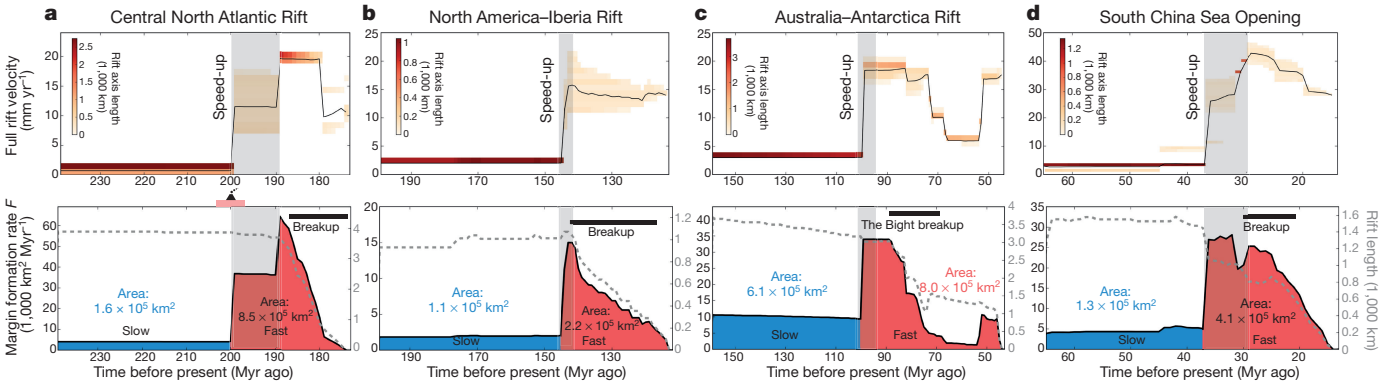


Figure 2 | Other rift systems. a–**d**, All depicted rifts exhibit a two-phase velocity history. A rapid plate acceleration precedes inception of breakup by ~10 Myr. Timing of Central Atlantic Magmatic Province is depicted by volcano symbol in **a**. Corresponding map views are shown in Extended Data Figs 1–3. For an animation of each rift kinematic evolution, see Supplementary Videos 2–5.

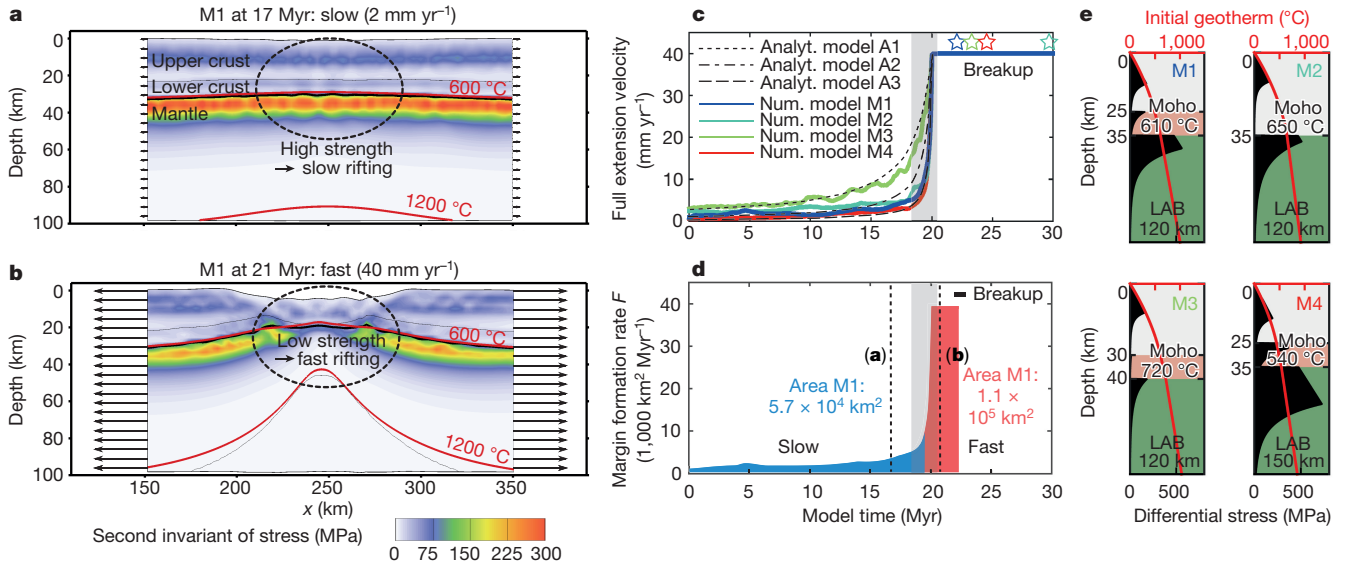


Figure 3 | Analytical and numerical model with force boundary conditions. a, **b**, Strength evolution of two-dimensional numerical rift model. Arrows at lateral model sides indicate extension velocity. **c**, Velocity evolution of analytical (Analyt.) models (A1–A3) and numerical (Num.) experiments (M1–M4). Analytical solution A1 employs a stress-exponent of $n = 3.5$, assuming an entirely viscous lithosphere, while A2 and A3 also account for brittle crustal deformation by using $n = 10$ and $n = 30$, respectively. Numerical reference model M1 is described in Methods and Extended Data Table 1. Other models are identical to M1, but apply a fully felsic crust (M2), a comparably thick crust (M3), or a thick lithosphere (M4), respectively. The duration of plate speed-up is a few million years for any of these configurations. The generated margin structures of M1–M4 are displayed along with six alternative models in Extended Data Fig. 7. **d**, The margin area of M1 consists of two major parts, formed during the slow and the fast phase, respectively. A constant rift length of 1,000 km is assumed. **e**, Rheological setup of models M1–M4. LAB, lithosphere–asthenosphere boundary; Moho, crust–mantle boundary.

and causes conjugate rift sides to accelerate rapidly[15,16]. We show ten models with varying rheological flow laws, thermal configurations, layer thicknesses, frictional softening and thermal expansivity (Fig. 3), which reproduce a large variety of rifted margin configurations (Extended Data Fig. 7). The two-phase behaviour and the abruptness of speed-up are robustly represented by any of these models. The numerical experiments also demonstrate that a variation of the extensional tectonic forces applied at the model boundaries affects the duration of the first, slow rift phase. Increasing the boundary force leads to earlier rift acceleration and breakup, while reducing the force prolongs the slow rift phase, or even generates a failed rift where conductive cooling and thermal strengthening decrease the extension rate until the rift becomes abandoned (Extended Data Fig. 7). We conclude that our plate kinematic and theoretical analyses independently suggest two-phase velocity behaviour as a key feature of successful rifts, which should have affected any rifted continental margin. Numerical modelling with realistic material properties as conducted here brackets the duration of rift-induced plate speed-up to between 2 and 10 million years.

Our results have profound implications for the interpretation of passive margin structures. A variety of studies illustrate striking differences between proximal and distal rifted margin domains[3,10]. This dichotomy can be attributed to the suggested two-phase velocities during basinward localization: the slow phase shapes the proximal margin, while the fast phase dominates the distal margin where our analysis predicts larger fault-slip rates, faster subsidence, higher heat flow, enhanced partial melting and associated underplating or volcanism.

This study provides an alternative explanation for the often enigmatic lack of extension interpreted from the tectono-stratigraphic record along many rifted margins[17]. Both reconstructions and modelling suggest that most of the new area of crust formed during rifting is created in a comparably short period of time towards the end of the syn-rift phase, when strain is likely to have localized to the distal part of the margins. This may explain why extension estimates from syn-rift faulting, typically biased towards the proximal margin areas, and interpreted using concepts borrowed from failed rifts where the feedback process we describe never occurred, commonly underestimate the total extension indicated by whole-crustal thinning.

Our results are applicable to any rifted margin, whether volcanic or magma-poor. Yet, the evolution of specific rifts will be modulated by other factors affecting the force budget and the rift strength such as lithospheric heterogeneity, rift obliquity[13], active rifting due to asthenospheric upwelling[18] and diking, as well as plume arrival[19]. Plumes contribute to breakup by reducing the strength of the rift[19], however, there can be a considerable delay between plume arrival and abrupt plate acceleration[20]. While any combination of these processes may influence how lithospheric strength evolves within individual rift systems, all successful rifts are expected to experience the proposed strength–velocity feedback before breakup.

Owing to the high viscosity of the lower mantle, only lithospheric and upper mantle processes can affect plate movements at timescales of a few million years[21]. It has been proposed that abrupt plate accelerations can be caused by plume–lithosphere interaction[22], subduction initiation[23], and slab detachment[24] possibly induced by ridge subduction[25]. However, none of these mechanisms explain our result that plate speed-up systematically precedes continental breakup. While the present motions of Earth's plates are governed by slab pull, basal drag and ridge push, we propose that abrupt plate acceleration during continental rifting is controlled by the rapid decrease of rift strength.

Dynamic rift weakening presents a new explanation for several previously unexplained rapid absolute plate motion changes, often appearing as cusps or kinks in apparent polar wander (APW) paths. A recent review[26] found four cusps in global APW paths during the last 200 Myr and each of them can be associated with rift velocity speed-up during major continental rifting events: (1) 190 Myr ago—central North Atlantic opening; (2) 150 Myr ago—separation of East and West Gondwana; (3) 125 Myr ago—South Atlantic Rift and the split between India and Antarctica; and (4) 50 Myr ago—northeast Atlantic opening. A global-scale plate reorganization at ~100 Myr ago (ref. 27) corresponds to an increase in rift velocity between Australia and Antarctica, and the end of a standstill in the APW path for South

America as the final continental connection with Africa is broken[8]. We suggest that absolute plate motion changes are strongly related to continental breakup, allowing a linkage between palaeomagnetic data and geological evidence in reconstructing the dynamics of previous supercontinents such as Pangea, Rodinia and Nuna.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 2 February; accepted 5 May 2016.
Published online 18 July 2016.

1. Burov, E. B. in *Treatise on Geophysics* Vol. 6 (ed. Schubert, G.) 99–151 (Elsevier, 2007).
2. Whitmarsh, R. B., Manatschal, G. & Minshull, T. A. Evolution of magma-poor continental margins from rifting to seafloor spreading. *Nature* **413,** 150–154 (2001).
3. Péron-Pinvidic, G. & Manatschal, G. The final rifting evolution at deep magma-poor passive margins from Iberia-Newfoundland: a new point of view. *Int. J. Earth Sci.* **98,** 1581–1597 (2009).
4. Pérez-Gussinyé, M., Morgan, J. P., Reston, T. J. & Ranero, C. R. The rift to drift transition at non-volcanic margins: insights from numerical modelling. *Earth Planet. Sci. Lett.* **244,** 458–473 (2006).
5. Brune, S., Heine, C., Perez-Gussinye, M. & Sobolev, S. V. Rift migration explains continental margin asymmetry and crustal hyper-extension. *Nature Commun.* **5,** 4014 (2014).
6. Seton, M. *et al.* Global continental and ocean basin reconstructions since 200 Ma. *Earth Sci. Rev.* **113,** 212–270 (2012).
7. Heine, C., Zoethout, J. & Müller, R. D. Kinematics of the South *Atlantic rift. Solid Earth* **4,** 215–253 (2013).
8. Granot, R. & Dyment, J. The Cretaceous opening of the South Atlantic Ocean. *Earth Planet. Sci. Lett.* **414,** 156–163 (2015).
9. Heine, C. & Brune, S. Oblique rifting of the Equatorial Atlantic: why there is no Saharan Atlantic Ocean. *Geology* **42,** 211–214 (2014).
10. Lavier, L. L. & Manatschal, G. A mechanism to thin the continental lithosphere at magma-poor margins. *Nature* **440,** 324–328 (2006).
11. Kusznir, N. J. & Park, R. G. The extensional strength of the continental lithosphere: its dependence on geothermal gradient, and crustal composition and thickness. *Geol. Soc. Lond. Spec. Publ.* **28,** 35–52 (1987).
12. Christensen, U. R. An Eulerian technique for thermomechanical modeling of lithospheric extension. *J. Geophys. Res. Solid Earth* **97,** 2015–2036 (1992).
13. Brune, S., Popov, A. A. & Sobolev, S. V. Modeling suggests that oblique extension facilitates rifting and continental breakup. *J. Geophys. Res.* **117,** B08402 (2012).
14. Bürgmann, R. & Dresen, G. Rheology of the lower crust and upper mantle: evidence from rock mechanics, geodesy, and field observations. *Annu. Rev. Earth Planet. Sci.* **36,** 531–567 (2008).
15. Takeshita, T. & Yamaji, A. Acceleration of continental rifting due to a thermomechanical instability. *Tectonophysics* **181,** 307–320 (1990).
16. Hopper, J. R. & Buck, W. R. The initiation of rifting at constant tectonic force: role of diffusion creep. *J. Geophys. Res. Solid Earth* **98,** 16213–16221 (1993).
17. Reston, T. J. The structure, evolution and symmetry of the magma-poor rifted margins of the North and Central Atlantic: a synthesis. *Tectonophysics* **468,** 6–27 (2009).
18. Huismans, R. S., Podladchikov, Y. Y. & Cloetingh, S. Transition from passive to active rifting: relative importance of asthenospheric doming and passive extension of the lithosphere. *J. Geophys. Res. Solid Earth* **106,** 11271–11291 (2001).
19. Buiter, S. J. H. & Torsvik, T. H. A review of Wilson Cycle plate margins: a role for mantle plumes in continental breakup along sutures? *Gondwana Res.* **26,** 627–653 (2014).
20. Brune, S., Popov, A. A. & Sobolev, S. V. Quantifying the thermo-mechanical impact of plume arrival on continental breakup. *Tectonophysics* **604,** 51–59 (2013).
21. Iaffaldano, G. & Bunge, H.-P. Rapid plate motion variations through geological time: observations serving geodynamic interpretation. *Annu. Rev. Earth Planet. Sci.* **43,** 571–592 (2015).
22. Cande, S. C. & Stegman, D. R. Indian and African plate motions driven by the push force of the Reunion plume head. *Nature* **475,** 47–52 (2011).
23. Gurnis, M., Hall, C. & Lavier, L. Evolving force balance during incipient subduction. *Geochem. Geophys. Geosystems* **5,** Q07001 (2004).
24. Bercovici, D., Schubert, G. & Ricard, Y. Abrupt tectonics and rapid slab detachment with grain damage. *Proc. Natl Acad. Sci. USA* **112,** 1287–1291 (2015).
25. Seton, M. *et al.* Ridge subduction sparked reorganization of the Pacific plate-mantle system 60–50 million years ago. *Geophys. Res. Lett.* **42,** 1732–1740 (2015).
26. Torsvik, T. H., Muller, R. D., Van der Voo, R., Steinberger, B. & Gaina, C. Global plate motion frames: toward a unified model. *Rev. Geophys.* **46,** RG3004 (2008).
27. Matthews, K. J., Seton, M. & Müller, R. D. A global-scale plate reorganization event at 105–100 Ma. *Earth Planet. Sci. Lett.* **355–356,** 283–298 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements S.B. was funded by the Marie Curie International Outgoing Fellowship 326115, the German Research Foundation Priority Program 1375 SAMPLE, and the Helmholtz Young Investigators Group CRYSTALS. S.E.W., N.P.B. and R.D.M. were supported by Science and Industry Endowment Fund project RP 04-174 and Australian Research Council grant IH130200012. Simulations were performed on the cluster facilities of the German Research Centre for Geosciences. Figures were created using matplotlib, Tecplot and Matlab. We thank X. Qin and J. Cannon for their efforts developing the GPlates portal and pyGPlates infrastructure.

Author Contributions S.B. and S.E.W. conceived the plate tectonic analysis. S.B. designed and conducted the thermo-mechanical modelling. S.B., S.E.W. and N.P.B. developed the pyGPlates workflow. S.B., S.E.W. and R.D.M. discussed and integrated the results. The paper was written by S.B. with contributions from all authors.

Author Information The rift velocity database is accessible via an open-access virtual-globe web interface through http://portal.gplates.org/cesium/?view=rift_v. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.B. (sascha.brune@gfz-potsdam.de).

Reviewer Information *Nature* thanks S. Buiter and R. Granot for their contribution to the peer review of this work.

METHODS

Rift kinematics. Quantitative restoration of continents to their pre-rift configuration during Pangea breakup involves estimating the amount of syn-rift extension from present-day crustal thickness, and accounting for uncertainties in these estimates[28,29]. The time of onset of rifting between two plates is constrained by geological evidence such as the ages of oldest syn-rift sediments and rift-associated volcanism. The direction and rate of divergence during rifting can be reconstructed by careful consideration of a diverse range of geological indicators such as seismic tectono-stratigraphy, dating of exhumed and volcanic rocks dredged/drilled within continent–ocean transitions, and fitting constraints from sections of the plate boundaries beyond the rift zone[7,30,31]. Note that in Mesozoic/Cenozoic tectonic reconstructions, plate rotations have to be discretized, whereas typical stage lengths are 5–10 Myr or even longer as observations usually do not permit building plate kinematic models with smaller stages. Our analysis combines independently conducted reconstructions (Supplementary Table 1) that account for recent geological and geophysical data sets.

COBs. The definition of COBs contains considerable uncertainties for many margins—indeed, the definition of a COB as a sharp boundary is conceptually problematic, with interpretations of geophysical data highlighting the complex crustal architecture within the transition from continental to oceanic domains[2,32]. Regions of transition can be several tens of kilometres wide, with complexities that vary between margins closer to volcanic or non-volcanic end-member scenarios. Our starting point for defining COBs were the geometries defined by Seton *et al.*[6]. We modified these using a synthesis of COB interpretations compiled from published crustal-scale geophysical data sets (see Supplementary Table 2). These data are primarily derived from seismic refraction experiments, but interpretations of crustal structure based on seismic reflection and gravity modelling are also included for regions where refraction data are sparse. Additional seismic constraints come from the data set of Winterbourne *et al.*[33], who identified unequivocal oceanic crust adjacent to continental margins along seismic profiles including some industry data, which are otherwise unavailable. The synthesis of margin-perpendicular profiles gives us a series of tie points along each margin, with which our COBs must be broadly consistent. To define COB polylines, we must interpolate between these tie points, which we did guided by first-order trends in maps of gravity derivatives and magnetic anomalies[29]. However, for the specific purposes of this study, an important consideration is that we use the COBs to define the orientation of the rift. For this reason, we have used deliberately simplified COB geometries with orientations that represent the first-order trend of each rifted margin. Using these constraints, we generated alternative COB geometries to define a range of possible COB locations. Our preferred COB set lies relatively ocean-ward, so that it includes areas where the basement is interpreted to comprise exhumed mantle or seaward dipping reflectors, but not basement formed by sea-floor spreading processes. To test how sensitive our results are to our COB interpretation, we generated a second set of COB geometries, defining the extreme landward limit of basement that is not clearly continental crust (see Extended Data Figs 5 and 6). We stress that the COBs used in this study define an envelope of possible COB locations suitable for our sensitivity tests, and are not a natural replacement for 'best-fitting' COB locations interpreted in other studies.

Observational evidence for timing of rift initiation. Our results, and particularly the occurrence of a two-phase, 'slow–fast' pattern in reconstructions of continental rifting, is sensitive to the age assigned to the onset of rifting. The condition for the slow–fast trend to disappear would be if the rift onset ages in our reconstruction model are erroneously old, such that rifting began later and proceeded (from the same full-fit configuration) at a faster rate. Hence, to establish the robustness of the slow–fast trend, we summarize geological evidence for the minimum age at which rifting began in each of the rift systems illustrated in our study, as well as observational evidence for accelerations in rift velocity that lend weight to our kinematic reconstructions.

South Atlantic Rift. We model South Atlantic rifting using the reconstruction of Heine *et al.*[7], with onset of slow rifting at ~150 Myr ago followed by acceleration at ~125 Myr ago. An extensive study linking biostratigraphy, lithostratigraphy and timing of deformation within basins along both African and South Atlantic conjugate margins north of the Walvis Fracture Zone[34] indicates that rifting was established by Berriasian times (>140 Myr ago). South of the Walvis Fracture Zone, the main phase of rifting between South America and southern Africa probably began earlier, in the Late Jurassic, following widespread, isolated Triassic–Jurassic rift basin development within southern South America[35]. The timing of slow rift onset and subsequent acceleration are consistent with earlier reconstructions[36]. A reconstruction invoking later, post-Aptian age of breakup between salt basins on the Brazilian and Angolan conjugate margins[37] has been proposed, but detailed seismic imaging and drilling of syn-rift sediments within the rifted margins, combined with basin subsidence histories[38] argue against this later breakup scenario. Further supporting our model, recent interpretation of magnetic anomalies in the southern South Atlantic within crust formed during the Cretaceous Normal Superchron constrains changes in the rate and direction of plate motions during breakup[8], showing rapid acceleration of plate divergence during initial breakup in the southernmost South Atlantic, contemporaneous with the final stages of rifting further north.

Central North Atlantic Rift. Our base reconstruction follows the work of Kneller *et al.*[28], who assigned a rifting onset age of 240 Myr ago, and predicts a speed-up at ~200 Myr ago. The speed-up occurs around the time of Central Atlantic Magmatic Province (CAMP) volcanism, which represents an important time marker within the rift evolution. Stratigraphic evidence from syn-rift sediments within basins along the eastern margins of North America[39,40] and the conjugate Northwest African margins[41] shows that continental rifting was active for at least 25 Myr before CAMP magmatism. Rapid speed-up of rifting is evidenced by a rate of sediment accumulation within these basins that increases drastically within 1–5 Myr before 200 Myr ago[39].

North America–Iberia. Our reconstruction proceeds from onset of rifting at 200 Myr ago with a speed-up at ~145 Myr ago. Recent, alternative reconstructions[42] show a similar increase in extension velocities at the end of the Jurassic, but with a slightly earlier initiation of rifting (~203 Myr ago). Evidence for widespread rifting is recorded in basins along the Newfoundland and Iberian margins beginning in the late Triassic, but resulting in little crustal thinning; a second phase beginning in the late Jurassic led to marked thinning and breakup in the Early Cretaceous[3,43]. Sequential restoration[30] yields post-145 Myr ago extension velocities of 1–2 mm yr^{-1}, consistent with our reconstructions. The rift velocity before ~145 Myr ago depends on the assumed age of rift onset. Taking the latest possible onset of rifting, Oxfordian (~161 Myr ago), modelled extension rates remain fairly constant throughout rifting. However, any earlier onset of rifting (for example, Triassic–Early Jurassic) as indicated by evidence listed above, would result in slower initial rifting followed by a Late Jurassic acceleration, consistent with our reconstruction model. The slow–fast velocity evolution is further supported by a study[44] that uses previously unpublished seismic and borehole data to show continuous rifting in the basins along western Iberia beginning in the Triassic (>210 Myr ago), with three rift cycles, ending at 144 Myr ago. These authors find that the subsidence is relatively slow in the first two rift phases, then increases rapidly in rate during a rift climax in Late Oxfordian–Kimmeridgian times (~160–152 Myr ago) coinciding with the timing of our speed-up.

Australia–Antarctica. Our reconstruction places the onset of rifting between Australia and Antarctica in the late Jurassic (~165–155 Myr ago), with an increase in rift velocity around 100 Myr ago. The earliest rift-fill comprises Callovian–Early Berriasian sediments (>160 to ~140 Myr ago), constrained by palynological dating of samples from the Polda, Bremer and Eyre basin[45]. Detailed tectono-stratigraphic analysis[45] and sequential structural restoration of interpreted seismic sections[46] point towards slow rifting during the Jurassic–Early Cretaceous, followed by a rapid acceleration in crustal thinning and subsidence at the beginning of the Late Cretaceous (beginning around 102 or 93.5 Myr ago)[46,47]. Shortly thereafter, breakup begins in the westernmost part of the rift system (dated by exhumation fabrics and volcanics), propagating eastwards over tens of millions of years[48].

South China Sea. Reconstructions of the South China margins indicate rapid extension and breakup beginning in the late Eocene[49]. Earlier extension developed from ~60 Myr ago within a former Andean-style margin, while onset of slow extension before the late Eocene speed-up is recorded by minor volcanism in rift-related basins[50]. This is further supported by subsidence and strain-rate analyses of wells and stratigraphic sections for basins within the northern margin of the South China Sea[51].

Gulf of California. The speed-up in our reconstruction occurs around 12 Myr ago, corresponding to a phase of abruptly increased rift velocity and obliquity inferred from widespread structural markers[52]. Phases of continental extension before the mid-Miocene are recorded by tectono-stratigraphic relationships and dating of rift-related volcanics and plutons[53,54]. These data indicate more diffuse extension at significantly lower rates: Ferrari *et al.*[53] estimate that the relative motion between the conjugate margins of the Gulf of California proceeded at an average of 7.7 mm yr^{-1} from ~30–18 Myr ago, and 8.3 mm yr^{-1} from ~18–12 Myr ago, consistent with our computed values.

North America–Greenland. Initiation of rifting by ~140 Myr ago is substantiated by dating of rift-related volcanics and biostratigraphy[55]. Starting from ~120 Myr ago, rift basin sedimentation is evident throughout the Cretaceous[56]. Rifting and subsidence rates were slow until a rapid increase around 70–80 Myr ago[55,57], around the time of an increase in rift velocity and subsequent initiation of sea-floor spreading.

Northeast Atlantic. Our reconstruction of relative motion between Greenland and Eurasia before the oldest sea-floor spreading magnetic anomalies in the northeast Atlantic (C24, ~53 Myr ago[58]) incorporates plate circuit computations using

constraints from the rift systems between Greenland, North America and Eurasia. The reconstruction shows slow extension in the Jurassic to Early Cretaceous followed by tectonic quiescence or modest mid-Cretaceous extension[59,60], and a pronounced speed-up in the Late Cretaceous. The acceleration in Late Cretaceous rifting indicated by the reconstructions is more tightly constrained from spreading histories in the adjacent basins. The most important phase of rifting that ultimately led to breakup is constrained to the Latest Cretaceous–Early Paleocene. Skogseid[61] proposed that enhanced syn-rift deposition took place between 75 and 62 Myr ago based on tectono-stratigraphy and subsidence analysis from seismic and well data from the Vøring margin.

End-member models of the South Atlantic Rift. We performed our analysis on several end-member models for the South Atlantic Rift (Extended Data Fig. 8). These models differ in terms of the timing of final South Atlantic breakup, which is difficult to constrain due to the Cretaceous superchron. However, models that feature a late breakup[8,37,62] also show a late speed-up at 110 Myr ago. Models with an earlier breakup, however[7,36] depict an earlier speed-up at 120 Myr ago. In all cases, initially slow rift velocities are followed by large plate accelerations that precede the final breakup by 10–15 Myr.

Another major difference between these models is the plate-internal deformation within South America: the Heine *et al.* model[7] uses a largely intact South American plate where deformation occurs only along the border to Patagonia, while Moulin *et al.*[62] separate South America in eight individual plates. We apply our analysis under two premises: (1) assuming a three-plate scenario (West Africa, South Africa, South America) shown in the upper panel of rift velocity diagrams (Extended Data Fig. 8b), and (2) we account for South America-internal deformation in a four-plate analysis where the northern and the southern part of South America are evaluated independently within the two bottommost panels in Extended Data Fig. 8b.

We find that plate models, which feature large intra-plate deformation[36,37,62] display two distinct speed-up events: first the Southern plates accelerate several million years before breakup in the southern South Atlantic and second the northern plates accelerate before breakup in the equatorial Atlantic. Plate models with less internal deformation, such as the Heine *et al.* model, display only a minor acceleration before southern South Atlantic opening whereas the largest speed-up occurs at 120 Myr ago before equatorial Atlantic breakup. While the relative importance of each speed-up depends on the amount of South America-internal deformation, all models illustrate plate acceleration before breakup of the controlling rift segment.

Database. The entire rift velocity database is accessible via an open-access virtual-globe web interface through http://portal.gplates.org/cesium/?view=rift_v. This database contains the rift velocity history of any point at a major post-Pangea rifted margin. The velocity history can be visualized online and downloaded, lending itself as a source of tectonic boundary conditions for basin analysis software and geodynamic forward models.

Numerical model setup. We apply the finite element code SLIM3D[63] to solve the coupled system of conservation equations for momentum, thermal energy and constitutive equations. The reference model M1 consists of four distinct petrological layers: 25 km of felsic upper crust[64], 10 km of mafic lower crust[65], and a lithospheric mantle dominated by dry olivine rheology[66] that extends to a depth of 120 km. The weak asthenospheric material below 120 km depth is represented through wet (that is, 1,000 p.p.m. H/Si) olivine rheology[66]. The entire model comprises a rectangular domain of 150 km depth and 500 km width, with 2 km resolution.

We apply dynamic boundary conditions at the lateral model sides, such that during rifting, the boundary force is kept constant, allowing for self-consistent evolution of extensional velocities. This approach is feasible if the model domain represents a large region whose strength is a major component in the overall force balance of the involved plates. The constant boundary force is maintained in our model until extensional velocities reach typical sea-floor spreading rates. Hereafter, the low rift strength becomes neglectable in the force balance, and we use velocity boundary conditions with a rate of 40 mm yr^{-1}. At the top boundary we use a free surface while at the bottom side isostatic equilibrium is realized by means of the Winkler foundation, where in- and outflow of material is accounted for during re-meshing. Deformation is accommodated by elasto-visco-plastic rheology so that the model self-consistently reproduces diverse lithospheric-scale deformation processes such as faulting, flexure and lower crustal flow. Viscous flow occurs via two creep mechanisms: diffusion and dislocation creep. The Mohr–Coulomb failure model is implemented for brittle deformation.

The thermal state at the model start is a steady-state temperature distribution resulting from each layer's heat conductivity, radiogenic heat production and the following boundary conditions: (1) lateral boundaries are thermally isolated, (2) the temperature at the surface is 0 °C, (3) below the lithosphere asthenosphere boundary the temperature is set to 1,350 °C. To avoid rift localization at the model boundaries, a small thermal heterogeneity is introduced in the model centre. We introduce this heterogeneity of triangular shape and 20 km width by elevating the initial 1,350 °C isotherm up to 10 km before thermal equilibration[5]. All rheological and thermal parameters are given in Extended Data Table 1.

Analytical solution. Here we derive a transparent analytical solution for finite amplitude necking of a homogeneous viscous layer consisting of a power-law material. A horizontal layer of initial thickness D_0 is extended by a constant line force F that is applied parallel to the layer. In an incompressible, free layer the mean layer-parallel deviatoric stress τ is half of the total stress, F/D_0 (ref. 67), that is, $\tau = \frac{1}{2}F/D_0$. The deviatoric stress further relates to the strain rate through the power law $\dot{\varepsilon} = B\tau^n$ where the pre-exponential factor B and the stress exponent n are material parameters. Note that B is often considered to be temperature-dependent with $B = A \times \exp(-E/(RT))$. On the basis of these relations a characteristic viscosity $\eta_c = \tau/\dot{\varepsilon} = \tau^{(1-n)}/B$ and a characteristic time $t_c = \eta_c/\tau = 1/B \times (\frac{1}{2}F/D_0)^{-n}$ can be defined[68].

Owing to mass conservation and incompressible flow, the horizontal stretching of the layer has to be balanced by vertical thinning: $1/L \times \mathrm{d}L/\mathrm{d}t = -1/D \times \mathrm{d}D/\mathrm{d}t$, where L is the length of the layer. The resulting extensional velocity $v = \mathrm{d}L/\mathrm{d}t$ can thus be written as

$$v = -L/D \times \mathrm{d}D/\mathrm{d}t \quad (1)$$

We assume that the upper and lower boundaries are traction-free and that no depth-dependent thinning occurs. These assumptions allowed derivation of closed analytical solutions for necking instabilities in boudinage mechanics and slab detachment[68] involving the time-dependent layer thickness D that can be expressed as:

$$D = D_0(1 - t/t_r)^{(1/n)} \quad (2)$$

where the time until rupture t_r relates to the aforementioned characteristic time[68] t_c such that $t_r = t_c/n$.

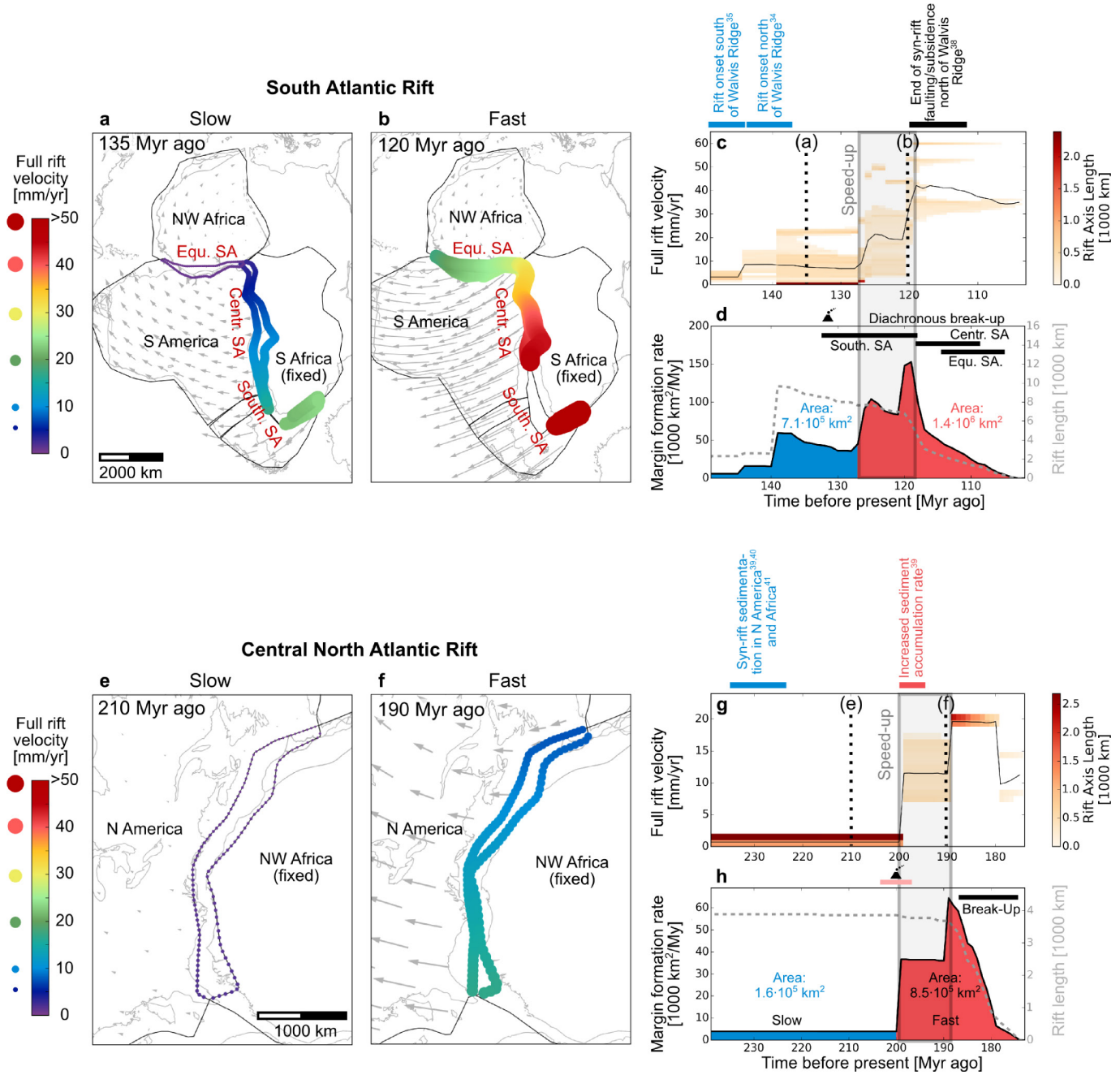
Combining equation (1) and equation (2) yields the formula for the time-dependent extension velocity:

$$v = L/(nt_r) \times (1 - t/t_r)^{-1} \quad (3)$$

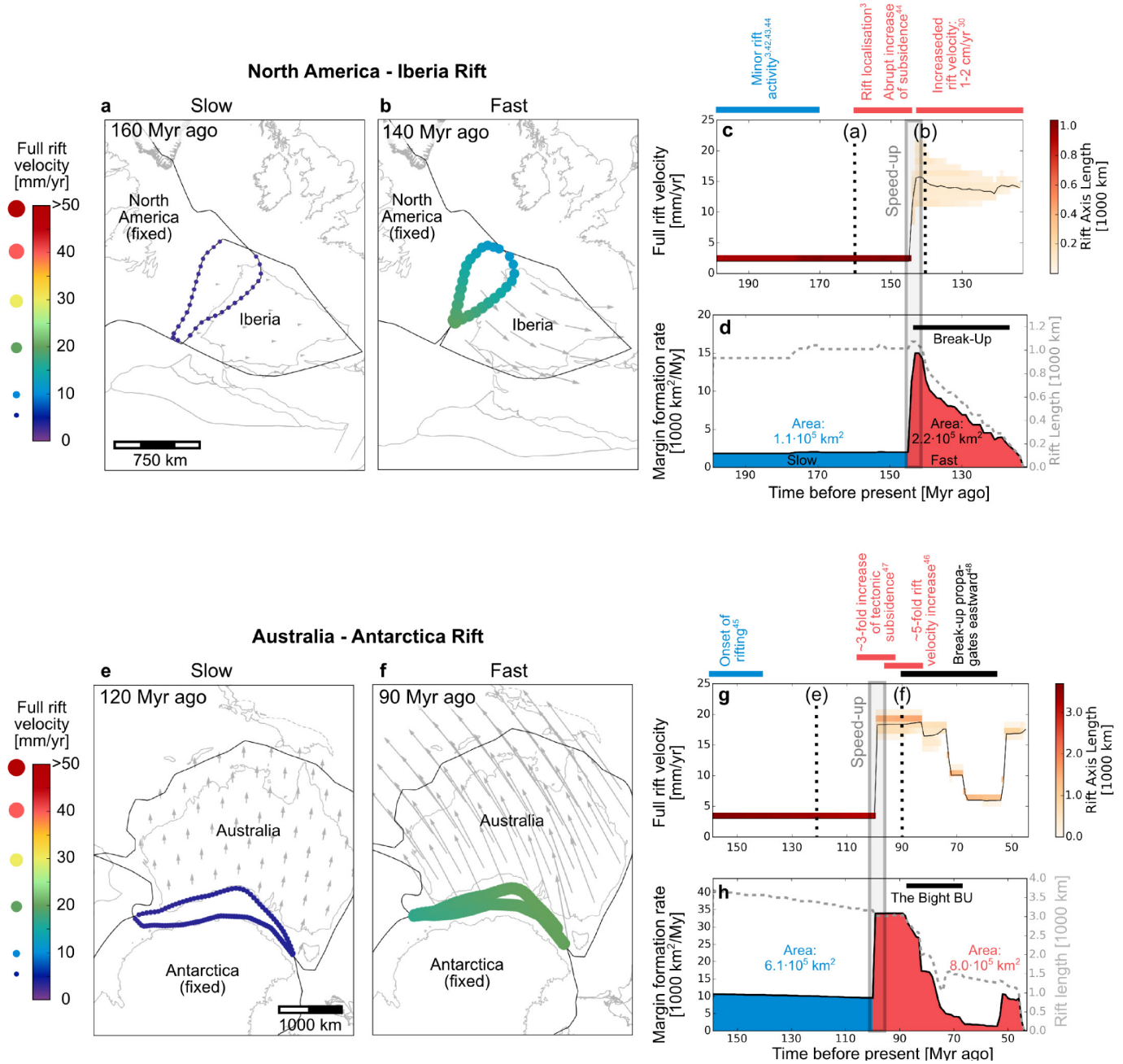
To apply the analytical results to continental rifting, we use parameters that describe a typical rift configuration (lithospheric thickness 100 km; mean lithospheric temperature $T = 600$ °C; applied tectonic force 8 TN m^{-1}; width of the necking zone $L = 100$ km; duration of rifting $t_r = 20$ Myr). The stress exponents of lithospheric materials range between 3 and 4, hence a purely viscous lithosphere can be approximated by $n = 3.5$. However, this approach neglects the existence of brittle deformation that is evidenced at real plate boundaries by ubiquitous faulting. Brittle failure can be represented as an end-member of power law creep, if stress exponents up to 30 are used[12]. The analytical solution is plotted in Fig. 3 for three cases: $n = 3.5$ (A1), $n = 10$ (A2), $n = 30$ (A3). Despite the simplicity of the analytical model, the numerical solutions of lithospheric necking are very similar to analytical solutions (Fig. 3). Hence, the analytical calculation corroborates our conclusion that it is the rapid loss of lithospheric strength during continental rifting, which is responsible for the abrupt increase of extension velocity.

28. Kneller, E. A., Johnson, C. A., Karner, G. D., Einhorn, J. & Queffelec, T. A. Inverse methods for modeling non-rigid plate kinematics: application to mesozoic plate reconstructions of the Central Atlantic. *Comput. Geosci.* **49,** 217–230 (2012).
29. Williams, S. E., Whittaker, J. M. & Müller, R. D. Full-fit, palinspastic reconstruction of the conjugate Australian-Antarctic margins. *Tectonics* **30,** TC6012 (2011).
30. Sutra, E., Manatschal, G., Mohn, G. & Unternehr, P. Quantification and restoration of extensional deformation along the Western Iberia and Newfoundland rifted margins. *Geochem. Geophys. Geosyst.* **14,** 2575–2597 (2013).
31. Whittaker, J. M., Williams, S. E. & Müller, R. D. Revised tectonic evolution of the Eastern Indian Ocean. *Geochem. Geophys. Geosyst.* **14,** 1891–1909 (2013).
32. Dean, S. L., Sawyer, D. S. & Morgan, J. K. Galicia Bank ocean–continent transition zone: new seismic reflection constraints. *Earth Planet. Sci. Lett.* **413,** 197–207 (2015).
33. Winterbourne, J., Crosby, A. & White, N. Depth, age and dynamic topography of oceanic lithosphere beneath heavily sedimented Atlantic margins. *Earth Planet. Sci. Lett.* **287,** 137–151 (2009).
34. Chaboureau, A.-C. *et al.* Paleogeographic evolution of the central segment of the South Atlantic during Early Cretaceous times: paleotopographic and geodynamic implications. *Tectonophysics* **604,** 191–223 (2013).
35. Loegering, M. J. *et al.* Tectonic evolution of the Colorado Basin, offshore Argentina, inferred from seismo-stratigraphy and depositional rates analysis. *Tectonophysics* **604,** 245–263 (2013).
36. Nürnberg, D. & Müller, R. D. The tectonic evolution of the South-Atlantic from Late Jurassic to present. *Tectonophysics* **191,** 27–53 (1991).

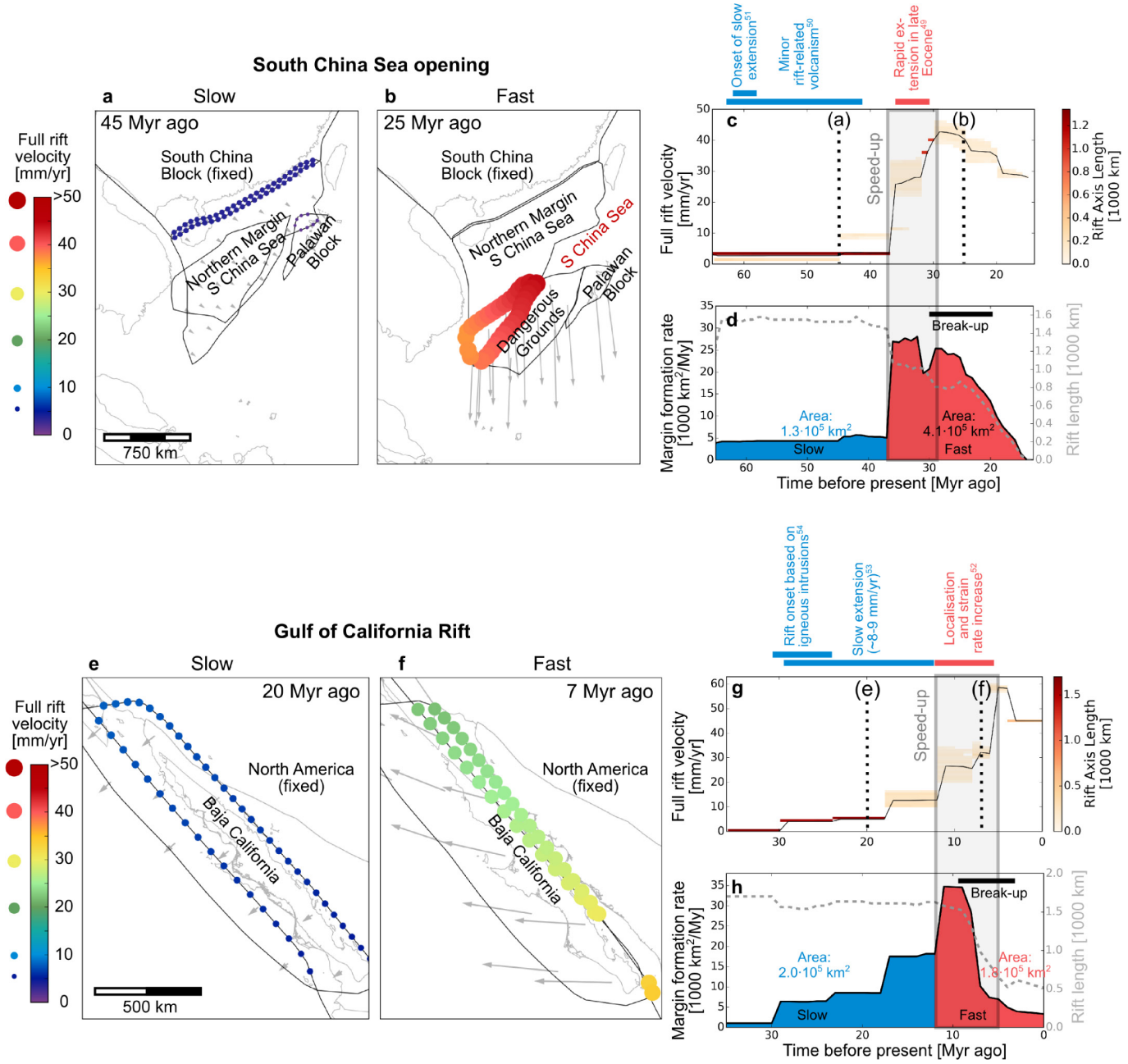
37. Torsvik, T. H., Rousse, S., Labails, C. & Smethurst, M. A. A new scheme for the opening of the South Atlantic Ocean and the dissection of an Aptian salt basin. *Geophys. J. Int.* **177,** 1315–1333 (2009).
38. Quirk, D. G. *et al.* Rifting, subsidence and continental breakup above a mantle plume in the central South Atlantic. *Geol. Soc. Lond. Spec. Publ.* **369,** 185–214 (2013).
39. Schlische, R. W., Withjack, M. O. & Olsen, P. E. Relative timing of CAMP, rifting, continental breakup, and basin inversion: tectonic significance. *Geophys. Monogr.* **136,** 33–59 (2003).
40. Withjack, M. O., Schlische, R. W., Malinconico, M. L. & Olsen, P. E. Rift-basin development: lessons from the Triassic–Jurassic Newark Basin of eastern North America. *Geol. Soc. Lond. Spec. Publ.* **369,** 301–321 (2013).
41. Davison, I. Central Atlantic margin basins of North West Africa: geology and hydrocarbon potential (Morocco to Guinea). *J. Afr. Earth Sci.* **43,** 254–274 (2005).
42. Vissers, R. L. M., van Hinsbergen, D. J. J., Meijer, P. T. & Piccardo, G. B. Kinematics of Jurassic ultra-slow spreading in the Piemonte Ligurian ocean. *Earth Planet. Sci. Lett.* **380,** 138–150 (2013).
43. Alves, T. M. *et al.* Diachronous evolution of Late Jurassic–Cretaceous continental rifting in the northeast Atlantic (west Iberian margin). *Tectonics* **28,** TC4003 (2009).
44. Pereira, R. & Alves, T. M. Tectono-stratigraphic signature of multiphased rifting on divergent margins (deep-offshore southwest Iberia, North Atlantic). *Tectonics* **31,** TC4001 (2012).
45. Ball, P., Eagles, G., Ebinger, C., McClay, K. & Totterdell, J. The spatial and temporal evolution of strain during the separation of Australia and Antarctica. *Geochem. Geophys. Geosyst.* **14,** 2771–2799 (2013).
46. Espurt, N. *et al.* Transition from symmetry to asymmetry during continental rifting: an example from the Bight Basin–Terre Adélie (Australian and Antarctic conjugate margins). *Terra Nova* **24,** 167–180 (2012).
47. Veevers, J. J. Change of tectono-stratigraphic regime in the Australian plate during the 99 Ma (mid-Cretaceous) and 43 Ma (mid-Eocene) swerves of the Pacific. *Geology* **28,** 47–50 (2000).
48. Direen, N. G., Stagg, H. M. J., Symonds, P. A. & Norton, I. O. Variations in rift symmetry: cautionary examples from the Southern Rift System (Australia–Antarctica). *Geol. Soc. Lond. Spec. Publ.* **369,** 453–475 (2013).
49. Lee, T.-Y. & Lawver, L. A. Cenozoic plate reconstruction of the South China Sea region. *Tectonophysics* **235,** 149–180 (1994).
50. Yan, Q., Shi, X. & Castillo, P. R. The late Mesozoic–Cenozoic tectonic evolution of the South China Sea: a petrologic perspective. *J. Asian Earth Sci.* **85,** 178–201 (2014).
51. Xie, H. *et al.* Cenozoic tectonic subsidence in deepwater sags in the Pearl River Mouth Basin, Northern South China Sea. *Tectonophysics* **615–616,** 182–198 (2014).
52. Bennett, S. E. K. & Oskin, M. E. Oblique rifting ruptures continents: example from the Gulf of California shear zone. *Geology* **42,** 215–218 (2014).
53. Ferrari, L. *et al.* Late Oligocene to Middle Miocene rifting and synextensional magmatism in the southwestern Sierra Madre Occidental, Mexico: the beginning of the Gulf of California rift. *Geosphere* **9,** 1161–1200 (2013).
54. Duque-Trujillo, J. *et al.* Timing of rifting in the southern Gulf of California and its conjugate margins: insights from the plutonic record. *Geol. Soc. Am. Bull.* **127,** 702–736 (2015).
55. Dickie, K., Keen, C. E., Williams, G. L. & Dehler, S. A. Tectonostratigraphic evolution of the Labrador margin, Atlantic Canada. *Mar. Petrol. Geol.* **28,** 1663–1675 (2011).
56. Chalmers, J. A. & Pulvertaft, T. C. R. Development of the continental margins of the Labrador Sea: a review. *Geol. Soc. Lond. Spec. Publ.* **187,** 77–105 (2001).
57. McGregor, E. D., Nielsen, S. B., Stephenson, R. A. & Haggart, J. W. Basin evolution in the Davis Strait area (West Greenland and conjugate East Baffin/Labrador passive margins) from thermostratigraphic and subsidence modelling of well data: implications for tectonic evolution and petroleum systems. *Bull. Can. Petrol. Geol.* **62,** 311–329 (2014).
58. Gaina, C., Gernigon, L. & Ball, P. Palaeocene–recent plate boundaries in the NE Atlantic and the formation of the Jan Mayen microcontinent. *J. Geol. Soc. Lond.* **166,** 601–616 (2009).
59. Tsikalas, F., Faleide, J. I., Eldholm, O. & Wilson, J. Late Mesozoic–Cenozoic structural and stratigraphic correlations between the conjugate mid-Norway and NE Greenland continental margins. *Geol. Soc. Lond. Pet. Geol. Conf. Ser.* **6,** 785–801 (2005).
60. Færseth, R. B. & Lien, T. Cretaceous evolution in the Norwegian Sea—a period characterized by tectonic quiescence. *Mar. Petrol. Geol.* **19,** 1005–1027 (2002).
61. Skogseid, J. Dimensions of the Late Cretaceous-Paleocene Northeast Atlantic rift derived from Cenozoic subsidence. *Tectonophysics* **240,** 225–247 (1994).
62. Moulin, M., Aslanian, D. & Unternehr, P. A new starting point for the South and Equatorial Atlantic Ocean. *Earth Sci. Rev.* **98,** 1–37 (2010).
63. Popov, A. A. & Sobolev, S. V. SLIM3D: A tool for three-dimensional thermo mechanical modeling of lithospheric deformation with elasto-visco-plastic rheology. *Phys. Earth Planet. Inter.* **171,** 55–75 (2008).
64. Gleason, G. C. & Tullis, J. A flow law for dislocation creep of quartz aggregates determined with the molten-salt cell. *Tectonophysics* **247,** 1–23 (1995).
65. Rybacki, E. & Dresen, G. Dislocation and diffusion creep of synthetic anorthite aggregates. *J. Geophys. Res.* **105,** 26017–26036 (2000).
66. Hirth, G. & Kohlstedt, D. L. Rheology of the upper mantle and the mantle wedge: a view from the experimentalists. *Geophys. Monogr.* **138,** 83–105 (2003).
67. Turcotte, D. L. & Schubert, G. *Geodynamics* (Cambridge Univ. Press, 2002).
68. Schmalholz, S. M. A simple analytical solution for slab detachment. *Earth Planet. Sci. Lett.* **304,** 45–54 (2011).
69. Andersen, O. B., Knudsen, P. & Berry, P. A. M. The DNSC08GRA global marine gravity field from double retracked satellite altimetry. *J. Geodyn.* **84,** 191–199 (2010).
70. Ranalli, G. & Murphy, D. C. Rheological stratification of the lithosphere. *Tectonophysics* **132,** 281–295 (1987).



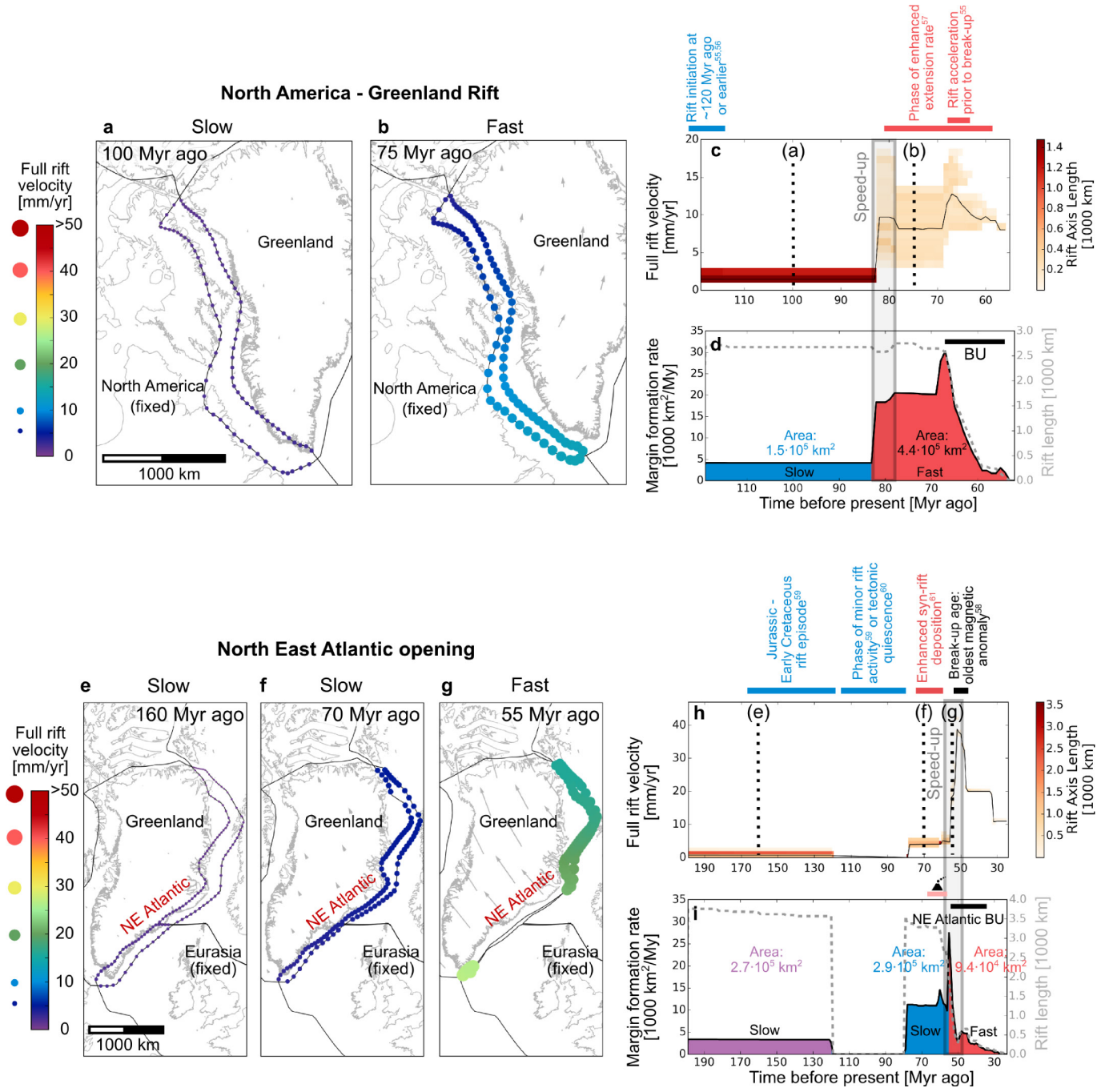
Extended Data Figure 1 | South Atlantic Rift and the central North Atlantic Rift. a–h, The maps depict snapshots of the slow and fast rift phase in the South Atlantic Rift (**a**, **b**) and central North Atlantic Rift (**e**, **f**). We corroborate the inferred velocity history with key temporal constraints[34,35,38–41] from geological and geophysical observations (**c**, **g**). For animations of the kinematic evolution, see Supplementary Videos 1 and 2.



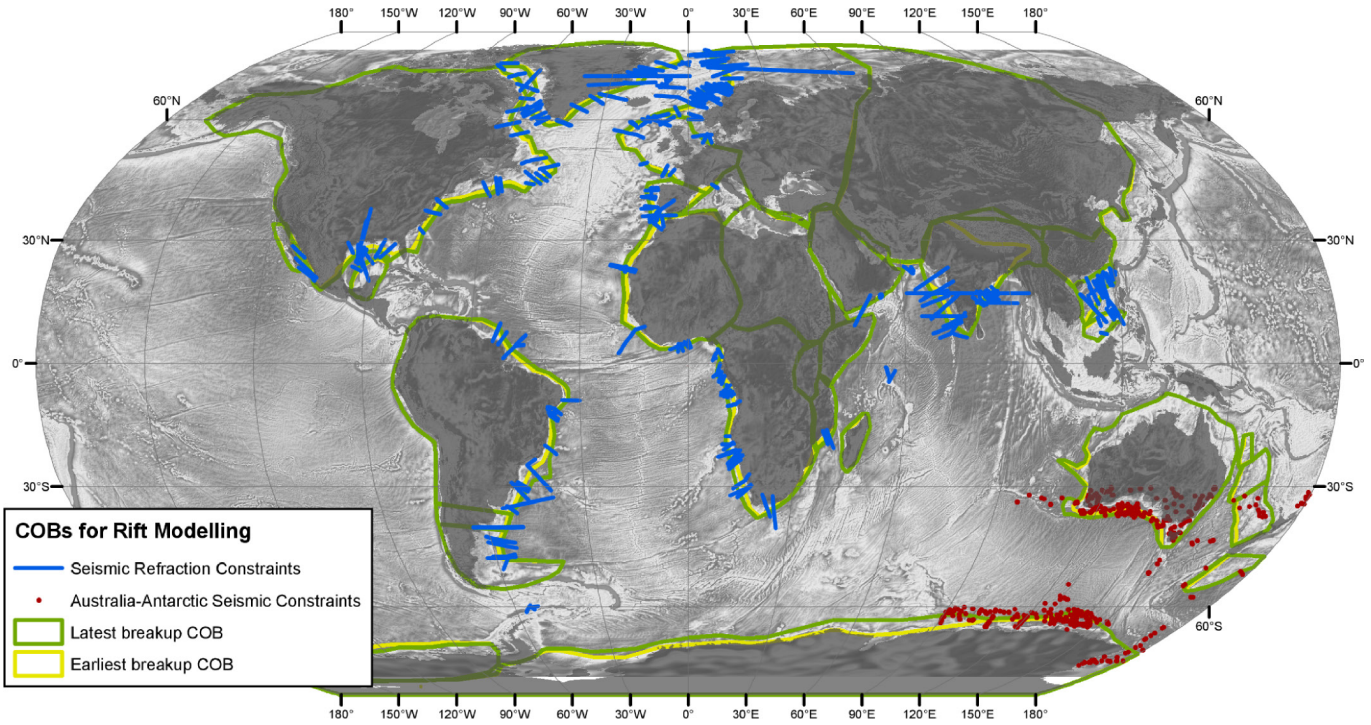
Extended Data Figure 2 | North America–Iberia Rift and the Australia–Antarctica Rift. a–h, The maps depict snapshots of the slow and fast rift phase in the North America–Iberia Rift (**a**, **b**) and the Australia–Antarctica Rift (**e**–**h**). We corroborate the inferred velocity history with key temporal constraints[3,30,42–48] from geological and geophysical observations (**c**, **g**). For animations of the kinematic evolution, see Supplementary Videos 3 and 4.



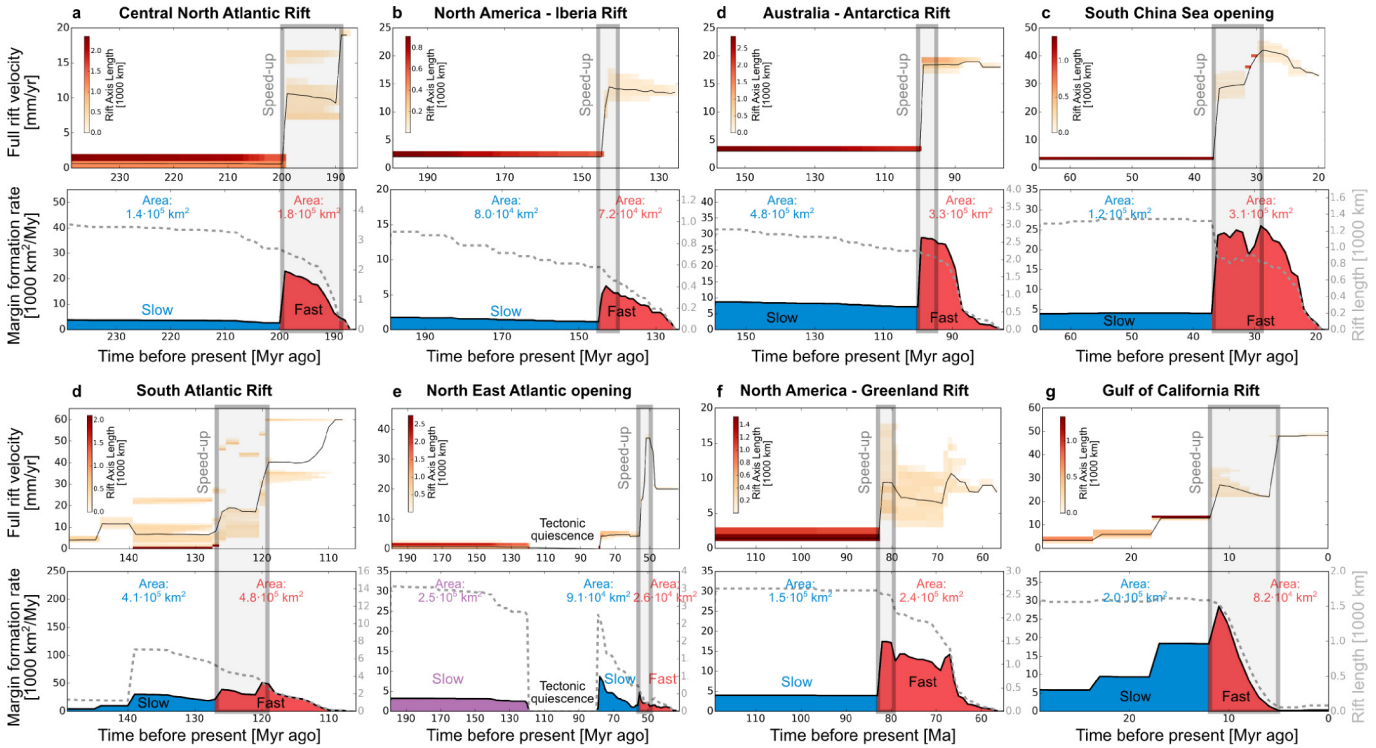
Extended Data Figure 3 | South China Sea opening and the Gulf of California Rift. a–h, The maps depict snapshots of the slow and fast rift phase in the South China Sea opening (**a**, **b**) and the Gulf of California Rift (**e**, **f**). We corroborate the inferred velocity history with key temporal constraints from[49–54] geological and geophysical observations (**c**, **g**). For animations of the kinematic evolution, see Supplementary Videos 5 and 6.



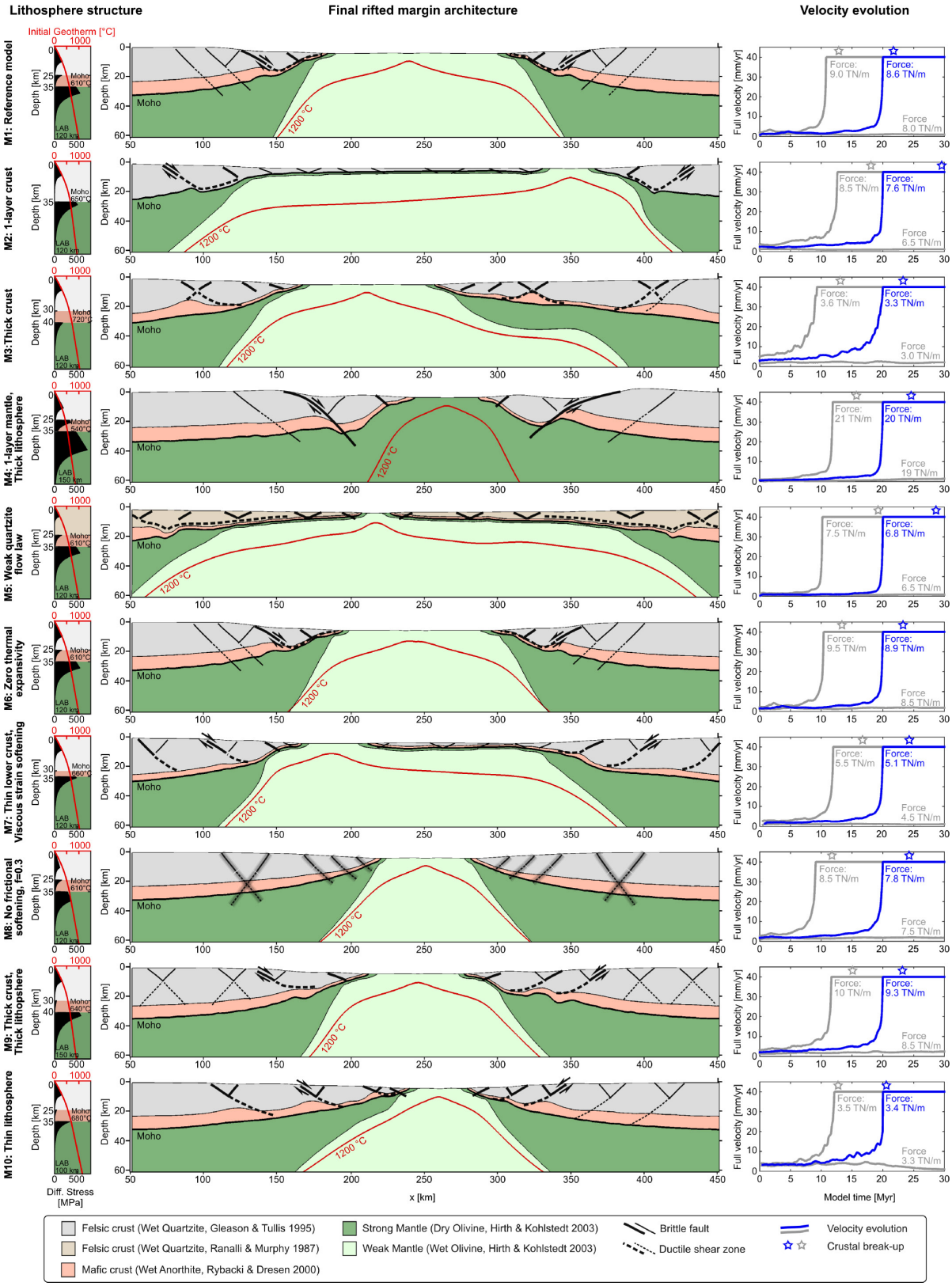
Extended Data Figure 4 | North America–Greenland Rift and the northeast Atlantic opening. a–i, The maps depict snapshots of the slow and fast rift phase in the North America–Greenland Rift (**a**, **b**) and the northeast Atlantic opening (**e**–**g**). We corroborate the inferred velocity history with key temporal constraints[55–61] from geological and geophysical observations (**c**, **h**). For animations of the kinematic evolution, see Supplementary Videos 7 and 8.



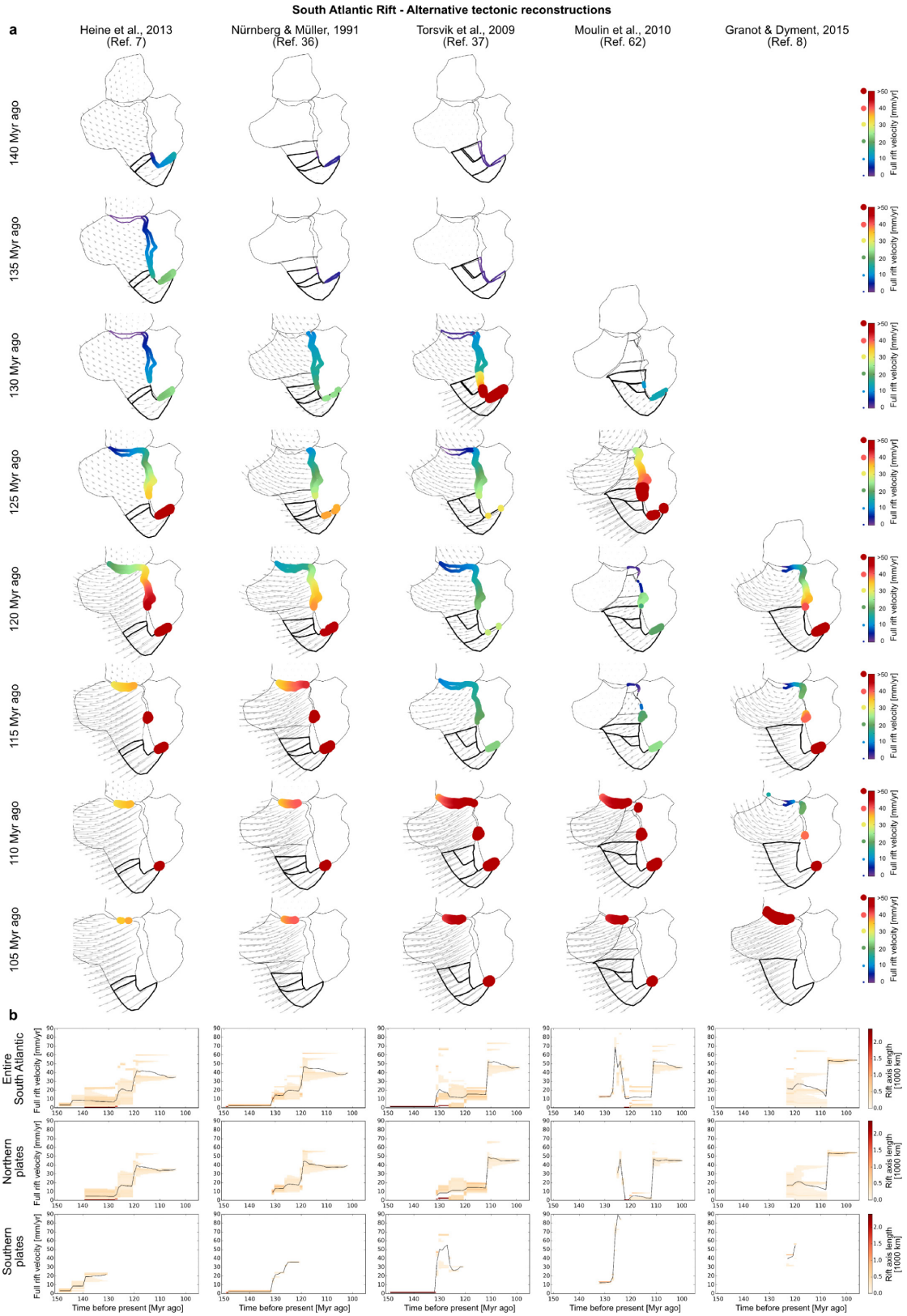
Extended Data Figure 5 | Data coverage for construction of COBs. We restrict our analysis to regions where seismic refraction data for both conjugate margins is available. Seismic refraction profiles are shown in blue, together with 'point' deep crustal seismic soundings linked together by gravity modelling. Red points represent a mixture of sonobuoys and deep reflection profiles. All references for displayed data are listed in Supplementary Table 2. Our preferred set of COBs (green) includes some areas where the basement is interpreted to comprise exhumed mantle or seaward dipping reflectors, but not basement formed by sea-floor spreading processes. The alternative set of COB geometries, defining the extreme landward limit of what basement that is not clearly continental crust, is shown in yellow. Underlying image shows global free-air gravity field[69].



Extended Data Figure 6 | Results using alternative, continent-ward set of COBs. The COB set is shown in Extended Data Fig. 5 as yellow polygons. Breakup takes place earlier, yet the two-phase evolution is robustly represented in this end-member scenario.



Extended Data Figure 7 | Final margin structures of numerical experiments. Using model M1 as our reference model, we vary layer thickness, rheological flow laws, the thermal configuration, frictional softening, and thermal expansivity to compute models M2–M10. M5 uses a comparatively weak quartzite flow law[70]. The final margin structures feature a wide range of rifted margin geometries reproducing all observed configurations of wide, narrow, symmetrical and asymmetrical margins. Depending on rheological evolution, extension is accommodated by brittle faults, ductile shear zones or both. For all cases, the associated time-dependent extension velocity (shown on the right in blue) exhibits the characteristic two-phase behaviour of slow rifting during the first rift phase, speed-up during lithospheric necking and fast rifting before breakup. Blue lines correspond to the final margin structures on the left and represent model runs where the boundary force coincides with the integrated strength of the yield strength profiles. Grey lines depict parameter variations where the boundary force is larger or smaller than the lithospheric strength resulting in two-phase velocities with an earlier speed-up, or decreasing rift activity reproducing failed rifts, respectively.



Extended Data Figure 8 | Alternative South Atlantic plate tectonic reconstructions. a, **b**, Several end-member models are shown that differ in terms of the timing of final South Atlantic breakup, and intra-plate deformation. **a**, Map view evolution. **b**, Frequency of extension velocity considering the entire South Atlantic (top) or only the Northern and Southern Plates (middle and bottom, respectively). Southern plates are depicted as bold polygons in the map view (**a**). Plate models[7,36] with a final breakup at ~110 Myr ago depict a speed-up at 125–120 Myr ago, while models with a later breakup[8,37,62] at ~100 Myr ago also involve a later rift acceleration at ~110 Myr ago. Reconstructions in which large intra-plate deformation[36,37,62] decouples northern and southern South America display first a speed-up of southern South America followed by a distinct speed-up of northern South America. Plate models with less internal deformation (for example, ref. 7) exhibit a minor acceleration of the southern plates followed by a large acceleration of entire South America. In all cases, plate kinematics show major speed-up about 10 Myr before breakup of the controlling rift segment.

Extended Data Table 1 | Thermo-mechanical reference parameters

Parameter	Units	Upper Crust		Lower Crust	Strong Mantle	Weak Mantle
Density	$kg\ m^{-3}$	2700		2850	3280	3300
Thermal expansivity	$10^{-5}\ K^{-1}$	2.7		2.7	3.0	3.0
Bulk modulus	GPa	55		63	122	122
Shear modulus	GPa	36		40	74	74
Heat capacity	$J\ kg^{-1}K^{-1}$	1200		1200	1200	1200
Heat conductivity	$W\ K^{-1}m^{-1}$	2.5		2.5	3.3	3.3
Radiogenic heat production	$\mu W\ m^{-3}$	1.5		0.2	0.0	0.0
Initial friction coefficient*	-	0.5		0.5	0.5	0.5
Cohesion	MPa	5.0		5.0	5.0	5.0
Rheology		Wet Quartzite	Wet Quartzite	Wet Anorthite	Dry Olivine	Wet Olivine
Flow law reference		64	70	65	66	66
Pre-exponential constant for diffusion creep	$Pa^{-1}s^{-1}$	-	-	-	2.25e-09	1.5e-09
Activation energy for diffusion creep	kJ/mol	-	-	-	375	335
Activation volume for diffusion creep	cm^{-3}/mol	-	-	-	6	4
Pre-exponential constant for dislocation creep	$Pa^{-n}s^{-1}$	8.57e-28	1.54e-17	1.79e-15	6.51e-16	2.12e-15
Power law exponent for dislocation creep	-	4.0	2.3	3.0	3.5	3.5
Activation energy for dislocation creep	kJ/mol	223	154	356	530	480
Activation volume for dislocation creep	cm^{-3}/mol	0	0	0	13	11

These listed parameters are used unless indicated otherwise. Pre-exponential constants of the flow laws[64–66,70] have been recalculated to account for flow laws written as a function of second invariants of stress and strain rate.
*During frictional strain softening, the friction coefficient reduces linearly from 0.5 to 0.05 for brittle strain between 0 and 1. For strains larger than 1, it remains constant at 0.05.

LETTER

doi:10.1038/nature19075

Emergence of a *Homo sapiens*-specific gene family and chromosome 16p11.2 CNV susceptibility

Xander Nuttle[1]*, Giuliana Giannuzzi[2]*, Michael H. Duyzend[1], Joshua G. Schraiber[1], Iñigo Narvaiza[3], Peter H. Sudmant[1]†, Osnat Penn[1], Giorgia Chiatante[4], Maika Malig[1], John Huddleston[1,5], Chris Benner[3], Francesca Camponeschi[6], Simone Ciofi-Baffoni[6,7], Holly A. F. Stessman[1], Maria C. N. Marchetto[3], Laura Denman[1], Lana Harshman[1], Carl Baker[1], Archana Raja[1,5], Kelsi Penewit[1], Nicolette Janke[1], W. Joyce Tang[8], Mario Ventura[4], Lucia Banci[6,7], Francesca Antonacci[4], Joshua M. Akey[1], Chris T. Amemiya[8], Fred H. Gage[3,9], Alexandre Reymond[2] & Evan E. Eichler[1,5]

Genetic differences that specify unique aspects of human evolution have typically been identified by comparative analyses between the genomes of humans and closely related primates[1], including more recently the genomes of archaic hominins[2,3]. Not all regions of the genome, however, are equally amenable to such study. Recurrent copy number variation (CNV) at chromosome 16p11.2 accounts for approximately 1% of cases of autism[4,5] and is mediated by a complex set of segmental duplications, many of which arose recently during human evolution. Here we reconstruct the evolutionary history of the locus and identify bolA family member 2 (*BOLA2*) as a gene duplicated exclusively in *Homo sapiens*. We estimate that a 95-kilobase-pair segment containing *BOLA2* duplicated across the critical region approximately 282 thousand years ago (ka), one of the latest among a series of genomic changes that dramatically restructured the locus during hominid evolution. All humans examined carried one or more copies of the duplication, which nearly fixed early in the human lineage—a pattern unlikely to have arisen so rapidly in the absence of selection ($P < 0.0097$). We show that the duplication of *BOLA2* led to a novel, human-specific in-frame fusion transcript and that *BOLA2* copy number correlates with both RNA expression ($r = 0.36$) and protein level ($r = 0.65$), with the greatest expression difference between human and chimpanzee in experimentally derived stem cells. Analyses of 152 patients carrying a chromosome 16p11.2 rearrangement show that more than 96% of breakpoints occur within the *H. sapiens*-specific duplication. In summary, the duplicative transposition of *BOLA2* at the root of the *H. sapiens* lineage about 282 ka simultaneously increased copy number of a gene associated with iron homeostasis and predisposed our species to recurrent rearrangements associated with disease.

To reconstruct the evolutionary history of the chromosome 16p11.2 region, we generated complete, reference-quality genome sequences[6] (Supplementary Table 1) for one orangutan, two chimpanzee and three human haplotypes (Fig. 1a and Extended Data Fig. 1). Comparison with mouse establishes the orangutan configuration as ancestral. In both humans and chimpanzees, the region has been independently restructured, nearly doubling in length primarily by the differential accumulation of segmental duplications (Fig. 1a and Extended Data Fig. 1a). We find six inversions have occurred in the African great apes within chromosome 16p11.2 (Extended Data Figs 2–4 and Supplementary Tables 2 and 3), a nonrandom clustering ($P < 1 \times 10^{-6}$), with breakpoints mapping near an ~20-kilobase-pair (kbp) low-copy repeat 16a (LCR16a) core duplicon. The core encodes a positively selected gene family (*NPIP*) that emerged on the human–African great ape lineage[7]. Only within the human lineage do large (>100 kbp) segmental duplications exist in a direct orientation flanking the autism critical region at breakpoint regions BP4 and BP5 (Extended Data Fig. 5a and Supplementary Table 4)[8], implying that susceptibility to large-scale CNV associated with disease[4,5,9] arose specifically within the human species.

Structural differences between human haplotypes are largely restricted to integral changes in the copy number of a 102-kbp block within both the proximal and distal breakpoint regions (Extended Data Fig. 1b). This block is composed of two different segmental duplications originating from chromosome 16: a 72-kbp segment duplicated from chromosome 16p12.1 carrying *NPIP* and a portion of the *SMG1* serine–threonine kinase gene (*SMG1P*) and a 30-kbp segment carrying three intact genes: *BOLA2*, *SLX1* and *SULT1A3* (Fig. 1a and Extended Data Fig. 1b). More than a dozen large-scale structural changes, including six duplicative transpositions (>830 kbp) from elsewhere on chromosome 16, are required to reconcile the organization of human and chimpanzee chromosome 16p11.2 (Extended Data Figs 3, 4 and Supplementary Table 3). Assuming a human–chimpanzee divergence time of 6 million years ago (Ma) (ref. 10) and a constant substitution rate, we estimate that a 95-kbp segment including *BOLA2* duplicated across the critical region ~282 ka (95% confidence interval 361–209 ka), around the time when *H. sapiens* emerged as a species[11] (Figs 1b and 2a, Extended Data Fig. 6 and Supplementary Tables 5–7).

We examined copy number diversity[12] of the duplicated genes mapping to the 102-kbp cassette—*BOLA2*, *SLX1* and *SULT1A3*—in humans, archaic humans and apes (Fig. 2b–c, Extended Data Fig. 7 and Supplementary Tables 8–10). We found that *BOLA2* is duplicated in all *H. sapiens* individuals examined, including archaic representatives of Neolithic and Mesolithic populations[13], as well as the oldest sequenced archaic human, Ust'-Ishim, estimated to have lived 45 ka (ref. 14). In sharp contrast, *BOLA2* is single copy (that is, diploid copy number = 2) in nonhuman primates and the archaic hominins Neanderthal[2] and Denisova[3] (Fig. 2b–c and Supplementary Table 8), consistent with our phylogenetic point estimate of the duplication age. Human genomes contain from three to eight diploid *BOLA2* copies, with at least one copy of the distal duplicate *BOLA2B* (range one to four copies; mean and median two) and at least two copies of the proximal ancestral *BOLA2A* (range two to five copies; mean and median four; Fig. 2c and Supplementary Table 8).

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA. [2]Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland. [3]Laboratory of Genetics, The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, California 92037, USA. [4]Dipartimento di Biologia, Università degli Studi di Bari 'Aldo Moro', Bari 70125, Italy. [5]Howard Hughes Medical Institute, Seattle, Washington 98195, USA. [6]Department of Chemistry, University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Florence, Italy. [7]Magnetic Resonance Center CERM, University of Florence, Via Luigi Sacconi 6, 50019, Sesto Fiorentino, Florence, Italy. [8]Benaroya Research Institute at Virginia Mason, Seattle, Washington 98101, USA. [9]Center for Academic Research and Training in Anthropogeny (CARTA), 9500 Gilman Drive, La Jolla, California 92093, USA. †Present address: Program in Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA.
*These authors contributed equally to this work.

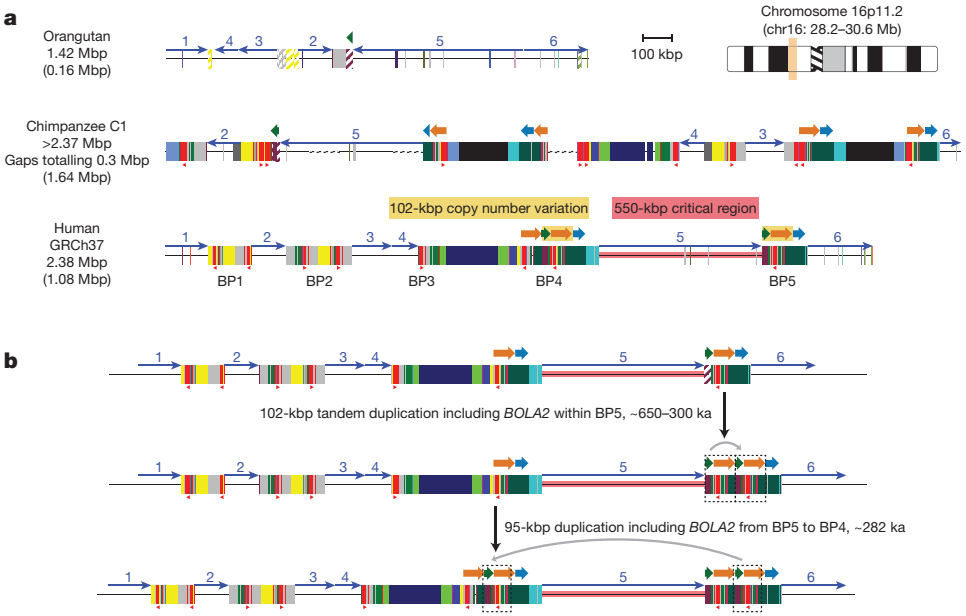


Figure 1 | Comparative sequence analysis of chromosome 16p11.2 among apes and the evolution of *BOLA2* duplications in humans. **a**, Genomic organization of chromosome 16p11.2 for one orangutan and one chimpanzee haplotype and the human reference haplotype (GRCh37 chr16:28195661–30573128). Blocks of segmental duplications within this locus mediate recurrent rearrangements in humans and have thus been defined as breakpoint regions BP1–BP5 (ref. 8). Coloured boxes and thick arrows indicate the extent and orientation of segmental duplications (different colours denote duplicons from different ancestral genomic loci; hashed boxes indicate sequence duplicated in humans but not in the species represented). Thin numbered arrows show orientations of gene-rich regions of unique sequence. Red triangles indicate locations and orientations of *NPIP* cores. Numbers (left) indicate the size of each haplotype, with the number of segmentally duplicated base pairs shown in parentheses. For chimpanzee, the size is a lower bound owing to gaps (dotted line sections) and the contig not reaching unique region 1. Regions of human CNV (yellow highlight) occur on both sides of the critical region and involve the same 102-kbp unit: a 30-kbp block (green arrow) containing *BOLA2*, *SLX1* and *SULT1A3* and a 72-kbp block (orange arrow) harbouring *SMG1P*. Expansion and contraction of this cassette underlie hundreds of kilobase pairs of structural diversity between human haplotypes. **b**, A model for the emergence of *BOLA2* duplications during *H. sapiens* evolution. It depicts structural changes over time leading to the present-day human architecture. A full evolutionary model detailing the dynamic evolution of chromosome 16p11.2 in great apes is provided in the Supplementary Information and Extended Data Figs 3 and 4.

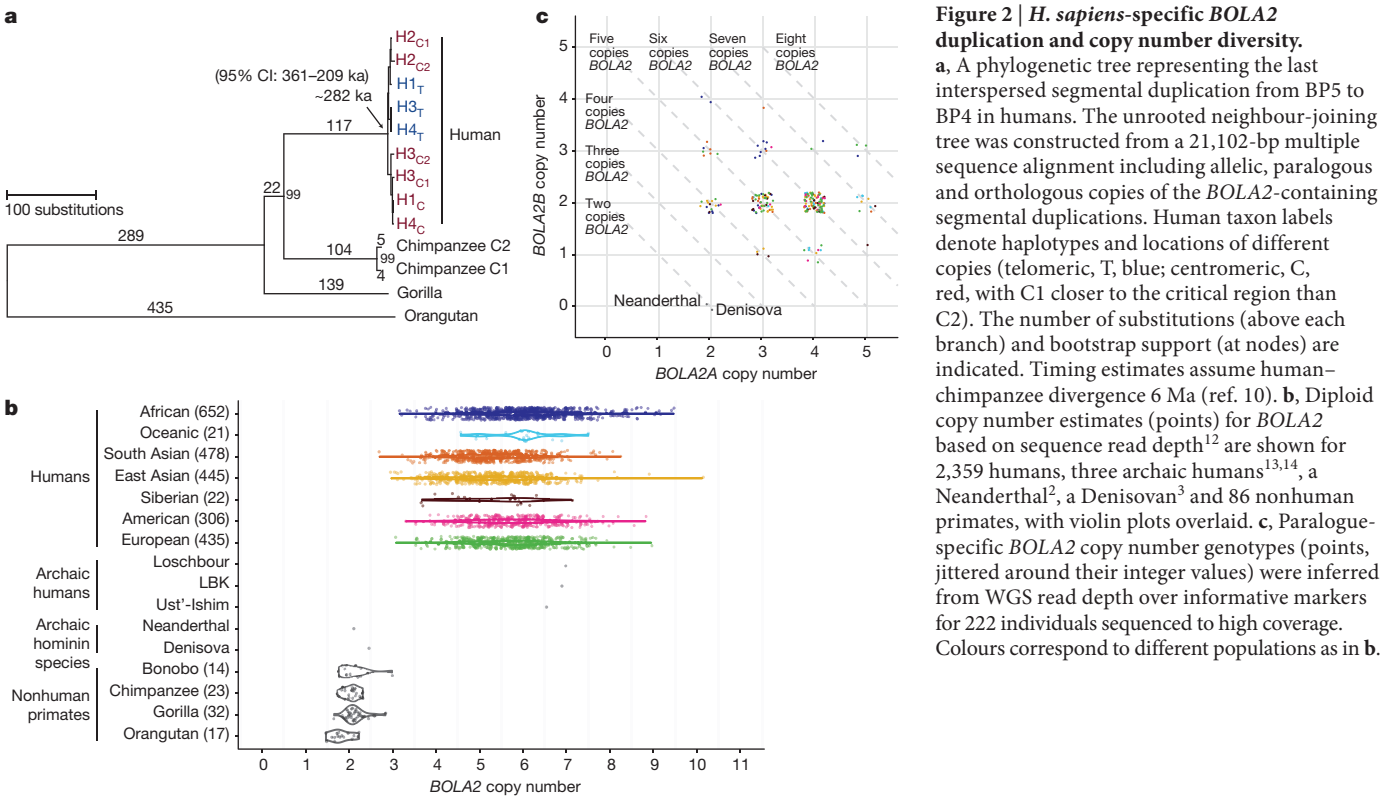


Figure 2 | *H. sapiens*-specific *BOLA2* duplication and copy number diversity. **a**, A phylogenetic tree representing the last interspersed segmental duplication from BP5 to BP4 in humans. The unrooted neighbour-joining tree was constructed from a 21,102-bp multiple sequence alignment including allelic, paralogous and orthologous copies of the *BOLA2*-containing segmental duplications. Human taxon labels denote haplotypes and locations of different copies (telomeric, T, blue; centromeric, C, red, with C1 closer to the critical region than C2). The number of substitutions (above each branch) and bootstrap support (at nodes) are indicated. Timing estimates assume human–chimpanzee divergence 6 Ma (ref. 10). **b**, Diploid copy number estimates (points) for *BOLA2* based on sequence read depth[12] are shown for 2,359 humans, three archaic humans[13,14], a Neanderthal[2], a Denisovan[3] and 86 nonhuman primates, with violin plots overlaid. **c**, Paralogue-specific *BOLA2* copy number genotypes (points, jittered around their integer values) were inferred from WGS read depth over informative markers for 222 individuals sequenced to high coverage. Colours correspond to different populations as in **b**.

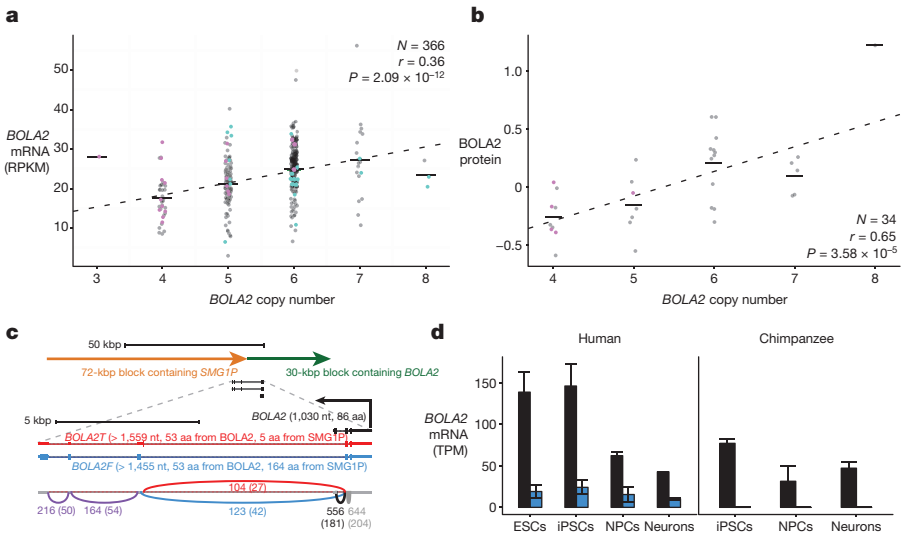


Figure 3 | *BOLA2* expression analyses. a, *BOLA2* mRNA expression quantifications[20] in 366 LCLs from individuals genotyped for *BOLA2* paralogue-specific copy number. Points indicate expression levels and copy number (jittered) for each cell line; horizontal lines show mean levels for each copy number. Dashed line shows least squares regression. Point colours indicate *BOLA2B* copy number (pink, one copy; black, two copies; cyan, three copies). Groups with the same aggregate *BOLA2* copy number but different combinations of paralogue-specific copy number do not exhibit differential expression, consistent with both paralogues producing mRNA. RPKM, reads per kilobase of transcript per million mapped reads. **b**, Plot layout as in **a**, but data show BOLA2 protein expression quantified by western blot densitometry on protein lysates from 34 LCLs. No evidence indicates differential protein expression of distinct *BOLA2* paralogues. **c**, *BOLA2* gene models, predicted protein products and support from RNA-seq data from human iPSCs. RT–PCR, cloning and capillary sequencing experiments identified three *BOLA2* isoforms: the canonical isoform (*BOLA2*, black) and two fusion isoforms consisting of the first two exons from canonical *BOLA2* fused with three exons from *SMG1P*. One fusion isoform (*BOLA2F*, blue) maintains the *BOLA2* open reading frame well beyond the fusion junction, whereas a third isoform (*BOLA2T*, red) contains a premature stop codon within the first *SMG1P*-derived exon. Numbers next to curved lines indicate mean counts of RNA-seq reads from two human iPSCs (two independent clones each) supporting each exon–exon junction, with standard errors in parentheses. nt, nucleotides; aa, amino acids. **d**, RNA-seq quantification of *BOLA2* (black) and *BOLA2F* (blue) mRNA expression through *in vitro* differentiation of primate iPSCs into neurons. Data from two human and two chimpanzee cell lines (two independent clones each, except for neurons) reveal higher levels of *BOLA2* transcripts in human iPSCs than in chimpanzee iPSCs and that *BOLA2* RNA levels decrease through neuronal differentiation. Bar heights indicate mean expression levels for each species and differentiation stage in transcripts per million (TPM); error bars, s.e.m. *BOLA2* expression in human embryonic stem cells (two cell lines) is consistent with data from human iPSCs. ESCs, embryonic stem cells; NPCs, neural progenitor cells.

In light of its recent origin and its potential to promote disease-causing rearrangement, we considered it remarkable that 99.8% of humans carry four or more copies of this segment. Ancient humans such as Ust'-Ishim as well as some of the oldest branches of modern humans (for example, San and Biaka pygmy[15]) typically carry five or six copies, indicating that it spread rapidly early in human history. We modelled various evolutionary scenarios by simulation on the basis of the observed genotypes and a realistic model of human demographic history (Extended Data Fig. 8a), assuming neutral evolution[16–18]. The observed genotypes or genotypes with higher *BOLA2B* frequencies only in humans were improbable ($P < 0.0097$; Extended Data Fig. 8b), even when the duplication age parameter was varied by an order of magnitude. Scenarios incorporating recurrent duplication were also deemed unlikely ($P < 0.0062$). We next implemented a model incorporating the 282 ka age estimate but varying the selection coefficient (s) as an input parameter, yielding a maximum likelihood estimate of $s = 0.0015$ (Extended Data Fig. 8c). Interestingly, the unique ~550-kbp critical region flanked by *BOLA2* duplications showed signatures consistent with a region under positive selection: the absence of archaic introgression[19], low diversity (bottom 2.7%) and an excess of rare variants (Extended Data Fig. 8d–e).

Because humans show extensive CNV, we assessed whether copy number correlated with messenger RNA (mRNA) and protein levels. We found a significant correlation between *BOLA2* copy number and expression at the RNA level from analysis of 366 lymphoblastoid cell lines (LCLs)[20] ($r = 0.36$, $P = 2.09 \times 10^{-12}$; Fig. 3a and Supplementary Tables 11 and 12) and at the protein level from analysis of whole-protein lysates from 34 LCLs ($r = 0.64$, $P = 4.34 \times 10^{-5}$; Fig. 3b and Supplementary Tables 13 and 14).

We also performed reverse transcription PCR (RT–PCR) and identified an alternative gene structure composed of the first two exons from *BOLA2* joined with three novel 3′ exons from an older segmental duplication containing *SMG1P* (Fig. 3c). This fusion isoform contains an open reading frame predicted to encode a 217-residue protein, including 53 residues from BOLA2 and 164 residues from SMG1P. Both canonical and fusion transcripts are co-expressed in a wide variety of tissues and developmental stages (Extended Data Fig. 9). Although the predicted fusion protein cannot be detected by existing antibodies, it is interesting that ribosome profiling data provide evidence that the mRNA is translated (Supplementary Table 15). Importantly, since the ancestral *BOLA2* at BP5 lacked the *SMG1P* duplication downstream, the origin of the fusion product must have coincided with the juxtaposition of *BOLA2* and *SMG1P* by the tandem 102-kbp segmental duplication ~650–300 ka at BP5. We conclude that this fusion isoform is *H. sapiens*-specific.

BOLA2 was previously identified as one of the top 50 genes differentially expressed between humans and nonhuman apes in induced pluripotent stem cells (iPSCs)[21], implying that this gene might be particularly relevant early in development. On the basis of our characterization of the different *BOLA2* isoforms, we revisited this observation by quantifying *BOLA2* mRNA levels by RNA sequencing (RNA-seq) in human and chimpanzee iPSCs, iPSC-derived neural progenitor cells and 8-week-old neurons. Remarkably, we found the greatest differences in canonical *BOLA2* expression at the iPSC state (twofold) and to a lesser extent in neural progenitor cells (1.5-fold) (Fig. 3d and Supplementary Table 16). Quantification of *BOLA2* expression in two primary human embryonic stem cell lines revealed transcript levels comparable to human iPSCs (Fig. 3d and Supplementary Table 16).

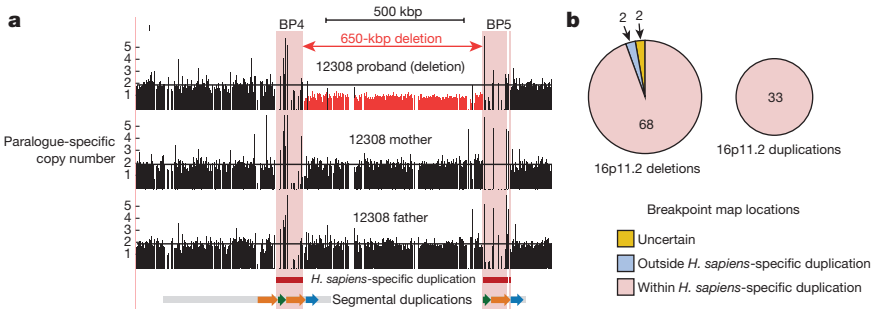


Figure 4 | Refinement of chromosome 16p11.2 rearrangement breakpoints. a, WGS results for a family with a *de novo* chromosome 16p11.2 microdeletion in a child with autism. Normalized read depth at unique 30-mer positions in the human reference genome GRCh37 is depicted for the proband, her mother and her father. Read-depth signatures reveal a deletion in the proband extending between but not beyond the *H. sapiens*-specific duplicated sequences (highlighted in pink). **b**, Summary of results across 105 independent microdeletion and microduplication events from 152 individuals; ~96% of breakpoints map to the *H. sapiens*-specific segmental duplication.

In contrast, examination of a panel of adult tissues[22] revealed no substantial differences in *BOLA2* mRNA levels between human and chimpanzee (Extended Data Fig. 9d). As expected, expression of the fusion *BOLA2–SMG1P* transcript was detected exclusively in human.

The duplication of *BOLA2* across the critical region expanded threefold the size of flanking high-identity, directly oriented sequence blocks (Extended Data Fig. 5a–b and Supplementary Tables 4, 17 and 18), theoretically predisposing the locus to recurrent CNV via unequal crossover (Extended Data Fig. 5c) specifically in the human lineage. To test this, we refined breakpoint locations in patients with autism and developmental delay carrying either the chromosome 16p11.2 microduplication or microdeletion event[23]. Using whole-genome sequence (WGS) data and a molecular inversion probe (MIP) assay[24], we localized breakpoints in 152 patients corresponding to 105 independent rearrangement events (Fig. 4a, Extended Data Fig. 10 and Supplementary Table 19). We found 96% (101 out of 105) of the disease-causing rearrangement breakpoints map within the *H. sapiens*-specific duplication containing *BOLA2* (Fig. 4b). Thus, the expansion of this segment rendered the chromosome 16p11.2 locus susceptible to recurrent rearrangement.

In summary, the level of genetic difference between humans and chimpanzees for chromosome 16p11.2 stands in sharp contrast to the oft-quoted 99% genetic identity between the species. The region has undergone extensive inversion and duplication, including a 95-kbp segment containing *BOLA2* that duplicated after our divergence with ancient hominins. This event contributes more derived sequence specific to *H. sapiens* than 35,500 previously reported human-specific single-nucleotide variants and indels combined[2]. The rapid rise and dispersal of this duplicated segment at the root of *H. sapiens* (~282 ka) are unlikely to have occurred under neutral evolution but rather are consistent with modest positive selection ($s = 0.0015$). The estimated strength of selection on the *BOLA2* duplication is an order of magnitude weaker than what is typically observed for recent positive selection (such as the emergence of lactase persistence ~10 ka (ref. 25)) but an order of magnitude stronger than nearly neutral mutations. Remarkably, the *BOLA2* duplication rapidly rose to high frequency in humans despite predisposing our species to recurrent CNV associated with disease. The expansion of this segment resulted in the formation of a novel fusion transcript and dramatic *BOLA2* expression differences between chimpanzee and human iPSCs. Although the phenotypic consequences of increased BOLA2 expression and the novel fusion transcript await future *in vivo* characterization, it is known that BOLA2 physically interacts in a heterotrimeric complex with GLRX3 (glutaredoxin 3)[26]. This complex is conserved from prokaryotes to humans[27] and was shown to have a role in iron sensing in yeast[28]. In vertebrates, BOLA2 has been hypothesized to play important roles in iron regulation[29] and iron–sulfur protein biogenesis[30]. We speculate that the expansion of this conserved gene may enhance iron utilization and homeostasis, especially during human embryonic development.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 29 October 2015; accepted 2 July 2016.
Published online 3 August 2016.

1. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188,** 107–116 (1975).
2. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505,** 43–49 (2014).
3. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338,** 222–226 (2012).
4. Weiss, L. A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **358,** 667–675 (2008).
5. Kumar, R. A. *et al.* Recurrent 16p11.2 microdeletions in autism. *Hum. Mol. Genet.* **17,** 628–638 (2008).
6. Huddleston, J. *et al.* Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res.* **24,** 688–696 (2014).
7. Johnson, M. E. *et al.* Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413,** 514–519 (2001).
8. Zufferey, F. *et al.* A 600 kb deletion syndrome at 16p11.2 leads to energy imbalance and neuropsychiatric disorders. *J. Med. Genet.* **49,** 660–668 (2012).
9. Jacquemont, S. *et al.* Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature* **478,** 97–102 (2011).
10. Patterson, N., Richter, D. J., Gnerre, S., Lander, E. S. & Reich, D. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441,** 1103–1108 (2006).
11. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
12. Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330,** 641–646 (2010).
13. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513,** 409–413 (2014).
14. Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514,** 445–449 (2014).
15. Tishkoff, S. A. *et al.* The genetic structure and history of Africans and African Americans. *Science* **324,** 1035–1044 (2009).
16. Yang, M. A., Harris, K. & Slatkin, M. The projection of a test genome onto a reference population and applications to humans and archaic hominins. *Genetics* **198,** 1655–1670 (2014).
17. Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18,** 337–338 (2002).
18. Ewing, G. & Hermisson, J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* **26,** 2064–2065 (2010).
19. Vernot, B. *et al.* Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* **352,** 235–239 (2016).
20. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511 (2013).
21. Marchetto, M. C. *et al.* Differential L1 regulation in pluripotent stem cells of humans and apes. *Nature* **503,** 525–529 (2013).
22. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478,** 343–348 (2011).
23. Simons VIP Consortium. Simons Variation in Individuals Project (Simons VIP): a genetics-first approach to studying autism spectrum and related neurodevelopmental disorders. *Neuron* **73,** 1063–1067 (2012).
24. Nuttle, X. *et al.* Rapid and accurate large-scale genotyping of duplicated genes and discovery of interlocus gene conversions. *Nature Methods* **10,** 903–909 (2013).
25. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74,** 1111–1120 (2004).

26. Li, H., Mapolelo, D. T., Randeniya, S., Johnson, M. K. & Outten, C. E. Human glutaredoxin 3 forms [2Fe-2S]-bridged complexes with human BolA2. *Biochemistry* **51,** 1687–1696 (2012).
27. Li, H. & Outten, C. E. Monothiol CGFS glutaredoxins and BolA-like proteins: [2Fe-2S] binding partners in iron homeostasis. *Biochemistry* **51,** 4377–4389 (2012).
28. Kumánovics, A. *et al.* Identification of FRA1 and FRA2 as genes involved in regulating the yeast iron regulon in response to decreased mitochondrial iron-sulfur cluster synthesis. *J. Biol. Chem.* **283,** 10276–10286 (2008).
29. Haunhorst, P. *et al.* Crucial function of vertebrate glutaredoxin 3 (PICOT) in iron homeostasis and hemoglobin maturation. *Mol. Biol. Cell* **24,** 1895–1903 (2013).
30. Banci, L., Camponeschi, F., Ciofi-Baffoni, S. & Muzzioli, R. Elucidating the molecular function of human BOLA2 in GRX3-dependent anamorsin maturation pathway. *J. Am. Chem. Soc.* **137,** 16133–16143 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank families at the participating Simons Variation in Individuals Project (Simons VIP) and Simons Simplex Collection sites, as well as the Simons VIP Consortium. Approved researchers can obtain the Simons VIP data set, the Simons Simplex Collection data set and/or biospecimens by applying at https://base.sfari.org. We thank M. Chaisson for single-molecule, real-time WGS data, B. Vernot for archaic introgression data, B. J. Nelson and K. Munson for technical assistance, M. L. Gage for editorial comments and T. Brown for assistance with manuscript preparation. This work was supported by the Paul G. Allen Foundation (grant 11631 to E.E.E.), the Simons Foundation Autism Research Initiative (SFARI 303241 to E.E.E. and 274424 to A.R.), the US National Institutes of Health (NIH grant 2R01HG002385 to E.E.E.), the Swiss National Science Foundation (31003A_160203 and CRSII33-133044 to A.R.) and funds from NIH TR01 MH095741, the Helmsley Charitable Fund, the Mathers Foundation and the JPB Foundation (to F.H.G.). X.N. was supported by a US National Science Foundation Graduate Research Fellowship under grant DGE-1256082. G.G. was awarded a Pro-Women Scholarship from the Faculty of Biology and Medicine, University of Lausanne. M.H.D. is supported by US National Institute of Mental Health grant 1F30MH105055-01. O.P. is a recipient of a Human Frontier Science Program postdoctoral fellowship. L.B. is supported by EC grant N653706, project iNEXT. S.C.B. and F.C. were supported by an Ente Cassa di Risparmio grant (2013/7201). E.E.E. is an investigator of the Howard Hughes Medical Institute. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author Contributions X.N., G.G., M.H.D., A.Re. and E.E.E. designed the study. X.N., G.G., M.H.D., M.M., J.H., L.D., L.H., C.Ba., A.Ra. and K.P. contributed to sequencing and assembly of haplotypes. X.N. developed the evolutionary model, with input from G.G. P.H.S. genotyped aggregate copy number from WGS data. X.N. and M.H.D. performed MIP experiments and analysed WGS data to genotype paralogue-specific copy number and refine rearrangement breakpoints. N.J. performed massively parallel sequencing. J.G.S., M.H.D. and X.N. performed population genetic simulations, with input from J.M.A. G.G. analysed RNA-seq data from LCLs, performed western blots and assessed the correlation of expression with copy number. I.N., C.Be. and M.C.N.M. performed and analysed RNA-seq experiments over *in vitro* differentiation of experimentally derived primate stem cells, with supervision from F.H.G. O.P., G.G. and X.N. analysed RNA-seq data from different human and nonhuman primate tissues. J.H. performed inversion density simulations using data provided by F.A. and M.V. G.C. and F.A. performed fluorescence *in situ* hybridization (FISH) experiments. F.C., S.C.B., H.A.F.S. and L.B. performed functional experiments and provided insights into potential effects of increased BOLA2 dosage. W.J.T. and C.T.A. constructed a bacterial artificial chromosome library. X.N. and E.E.E. wrote the paper, with input and approval from all co-authors.

Author Information Clone sequences, haplotype contig sequences and MIP data are available at the NCBI BioProject database under accession number PRJNA325679. RNA-seq data for neural progenitor cells and neurons are available at NCBI Gene Expression Omnibus under accession numbers GSE47626 and GSE83638. Patient WGS and MIP data are available at SFARI Base (https://sfari.org/resources/sfari-base) under accession numbers SFARI_SVIP_WGS_1 and SFARI_SVIP_MIPS_1. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.E.E. (eee@gs.washington.edu) or A.R. (alexandre.reymond@unil.ch).

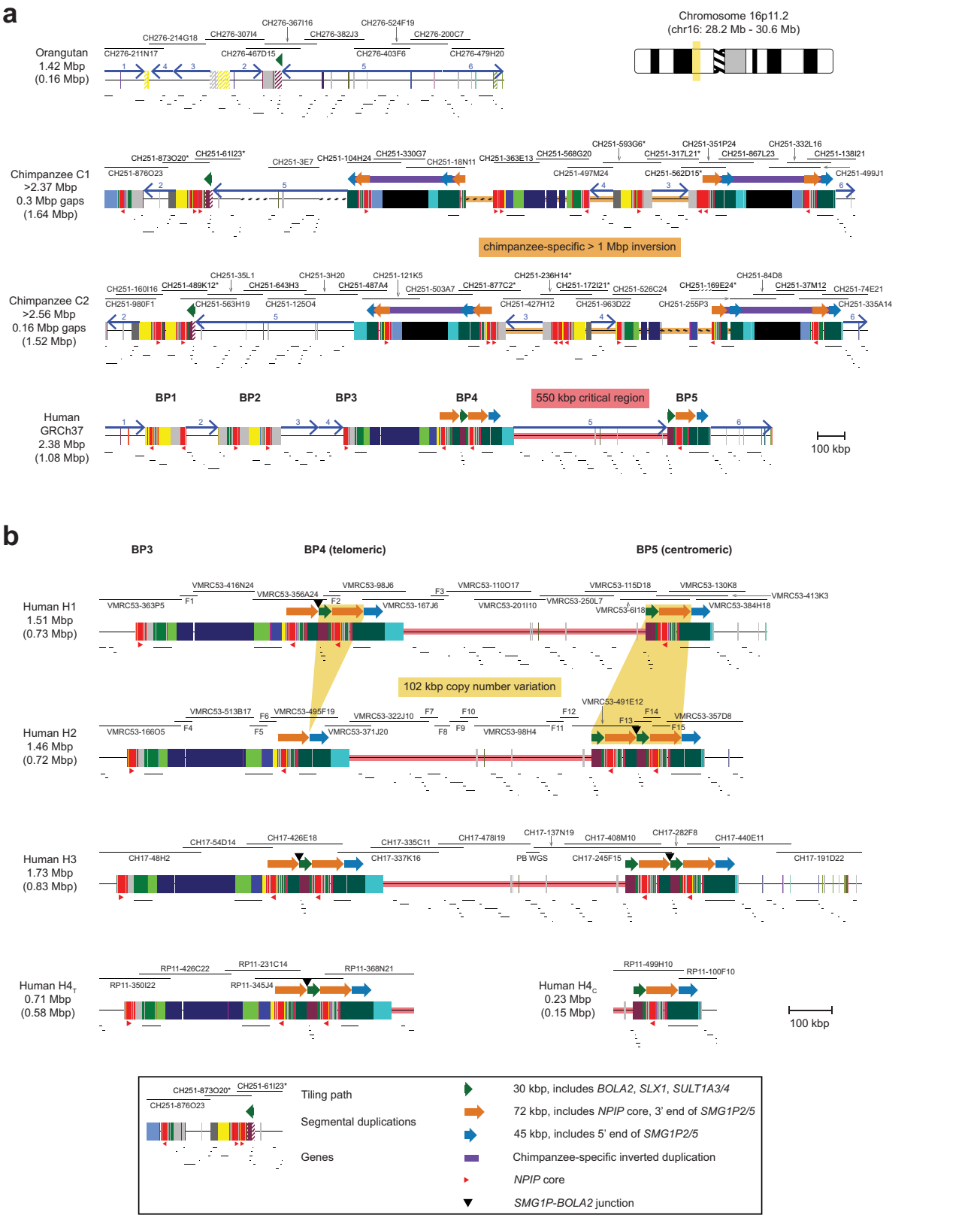
Reviewer Information *Nature* thanks D. Conrad, D. Haussler, C. Tyler-Smith and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

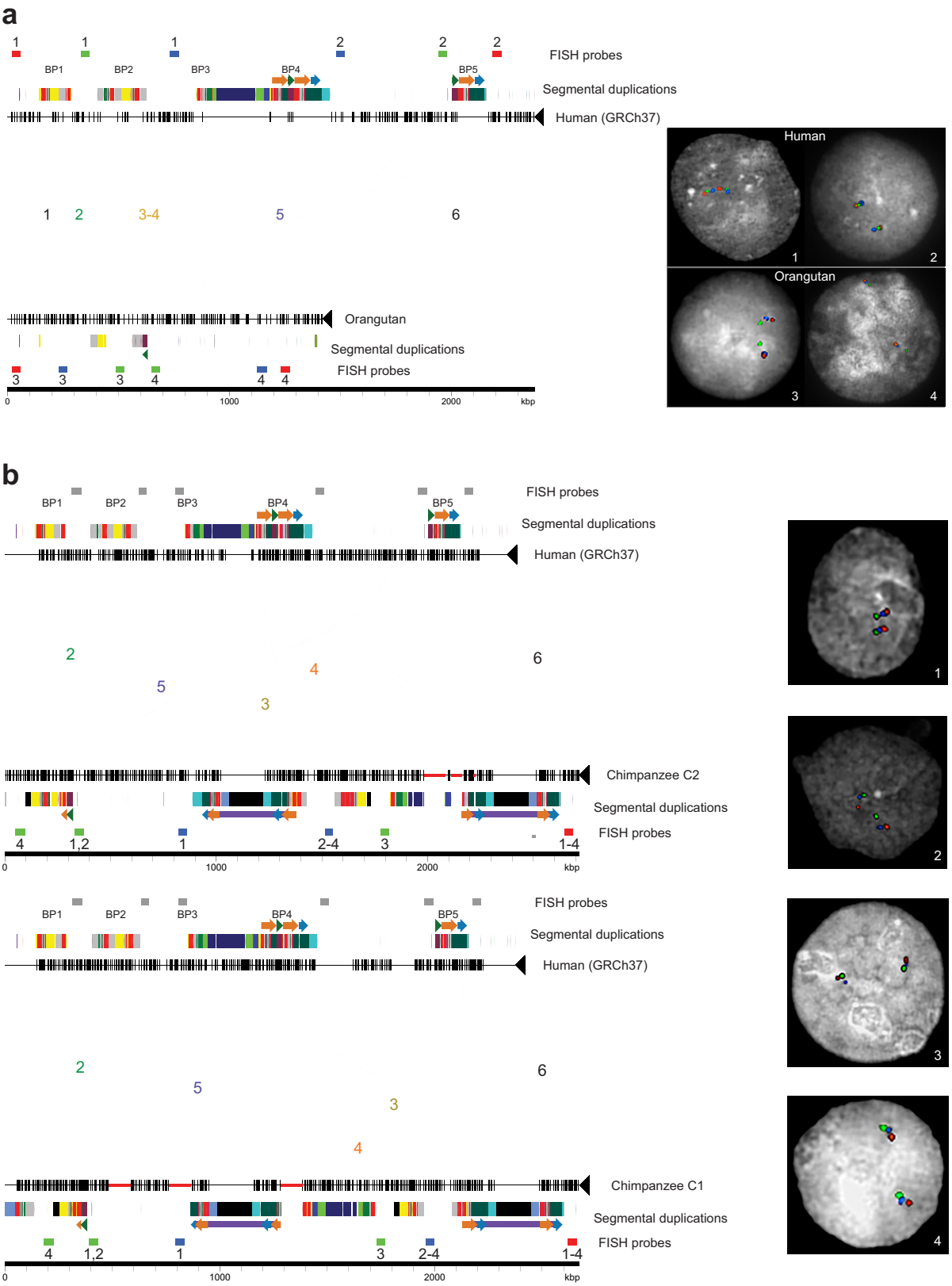
Single-molecule, real-time sequencing was used to generate high-quality sequence[6] from bacterial artificial chromosome clones obtained from genomic libraries. Clone sequences were assembled using HGAP and error-corrected using Quiver[31]. Contig assembly was performed using Sequencher (Gene Codes Corporation, Ann Arbor, Michigan) and validated by FISH. Copy number genotyping of genes and segmental duplications was performed using a read-depth method[12] and WGS data from humans[32,33], nonhuman primates[34] and archaic genomes[2,3,13,14], as well as single-molecule MIPs[35] targeted to paralogous sequence variants[24]. We estimated evolutionary timing of segmental duplication events on the basis of comparative sequencing and phylogenetic analyses (neighbour-joining method), adjusting branch lengths for trees that failed the Tajima's relative rate test and assuming divergence times of 6 Ma (human–chimpanzee)[10] and 15 Ma (human–orangutan). Evolutionary conservation analysis of *BOLA2* was performed by maximum likelihood (PAML). Likelihoods of *BOLA2B* fixation under different scenarios were assessed using the coalescent simulators *ms*[17] and *msms*[18], adapting a previously published demographic model[16]. *BOLA2* copy number estimates were correlated (Pearson's *r*) using RNA-seq quantifications[20] (PEER-normalized RPKM) and western blot BOLA2 densities in human LCLs grown in complete RPMI medium and lysed in RIPA buffer. After SDS–PAGE and transfer to PVDF membrane, blots were incubated with an anti-BOLA2 antibody (Santa Cruz Biotechnology, Dallas, Texas) and an anti-actin antibody (Sigma) for normalization purposes. Band densities were quantified using the Bio1D software. *BOLA2* coding DNA sequence (CDS) was cloned using the Gateway system (Invitrogen, Carlsbad, California). HeLa cells were transfected with cytomegalovirus-*BOLA2* CDS (both 10 and 17 kDa forms) and analysed by western blotting. *BOLA2* gene models were established via RT–PCR, cloning and capillary sequencing. RNA-seq data were generated from previously described embryonic stem cell and iPSC lines[21], as well as iPSC lines differentiated into neural progenitor cells and neurons. *BOLA2* mRNA expression was quantified in transcripts per million with Kallisto[36] (version 0.42.1) using a custom catalogue of transcripts including all human RefSeq transcripts with the three *BOLA2* isoforms. Breakpoints of chromosome 16p11.2 rearrangements were refined using Illumina whole-genome shotgun sequencing[37,38] and single-molecule MIP analysis[24,35,37] of patient DNA obtained from the Simons VIP[23] and Simons Simplex Collection[39]. All procedures for clinical assessment and blood extraction were approved by the institutional review boards of participating institutions, and informed consent was obtained for participation in this research.

31. Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods* **10,** 563–569 (2013).
32. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
33. Sudmant, P. H. *et al.* Global diversity, population stratification, and selection of human copy-number variation. *Science* **349,** aab3761 (2015).
34. Prado-Martinez, J. *et al.* Great ape genetic diversity and population history. *Nature* **499,** 471–475 (2013).
35. Hiatt, J. B., Pritchard, C. C., Salipante, S. J., O'Roak, B. J. & Shendure, J. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res.* **23,** 843–854 (2013).
36. Bray, N., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal RNA-seq quantification. Preprint at http://arxiv.org/abs/1505.02710 (2015).
37. Nuttle, X., Itsara, A., Shendure, J. & Eichler, E. E. Resolving genomic disorder-associated breakpoints within segmental DNA duplications using massively parallel sequencing. *Nature Protocols* **9,** 1496–1513 (2014).
38. Antonacci, F. *et al.* Palindromic *GOLGA8* core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability. *Nature Genet.* **46,** 1293–1302 (2014).
39. Fischbach, G. D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68,** 192–195 (2010).



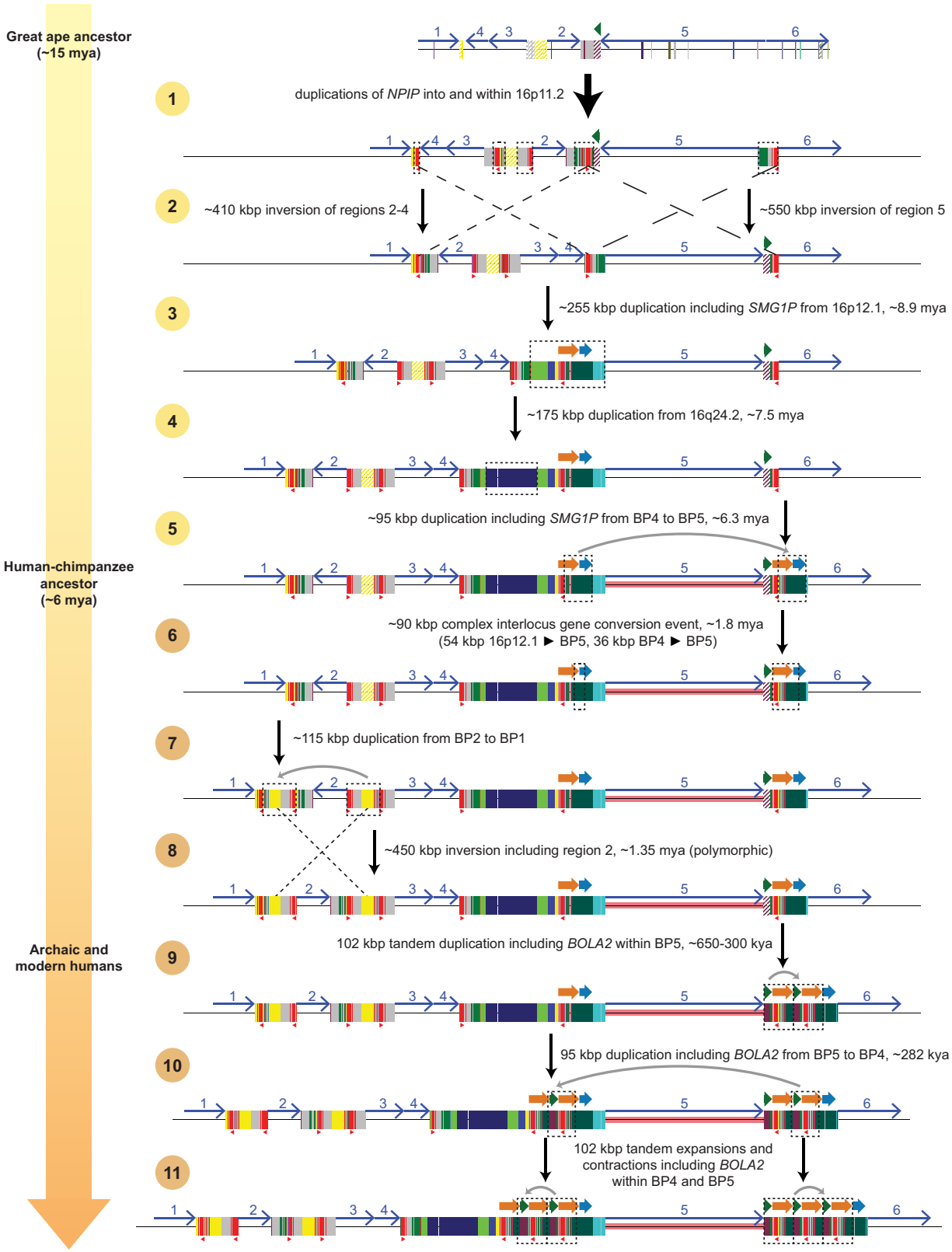
Extended Data Figure 1 | See next page for caption.

Extended Data Figure 1 | Comparative sequence analysis of chromosome 16p11.2 among apes. a, Genomic organization of chromosome 16p11.2 for one orangutan and two chimpanzee haplotypes and the human reference haplotype (GRCh37 chr16:28195661–30573128; see ideogram for approximate chromosomal location). Blocks of segmental duplications within this locus mediate recurrent rearrangements in humans; thus, these blocks have been defined as breakpoint regions BP1–BP5 (ref. 8). The ~550-kbp critical region (pink) and a >1-Mbp chimpanzee-specific inversion polymorphism (orange) are highlighted. Tiling paths of sequenced clones are indicated above each haplotype, with chimpanzee clones that could not be fully resolved marked with asterisks. Coloured boxes and thick arrows indicate the extent and orientation of segmental duplications (with different colours denoting duplicons from different ancestral genomic loci and hashed boxes indicating sequence duplicated in humans but not in the species represented). Thin numbered arrows show orientations of gene-rich regions of unique sequence. Numbers (left) indicate the size of each orthologous haplotype, with the number of segmentally duplicated base pairs shown in parentheses. Note that, for chimpanzee, these sizes are lower bounds owing to gaps in the contigs (dotted line sections) and the contigs not reaching unique sequence beyond BP1 (that is, unique region 1). **b**, Distinct human structural haplotypes over the chromosome 16p11.2 critical region and flanking sequences (three complete haplotypes extending from unique sequence distal to BP3 to unique sequence proximal to BP5 and one partial haplotype including BP3–BP4 and BP5 sequence contigs). High-quality sequence for each haplotype was generated by sequencing a total of 40 bacterial artificial chromosomes and 15 fosmids from three different human genomic libraries. Regions of CNV (highlighted in yellow along the first two haplotypes) occur on both sides of the critical region and involve the same 102-kbp unit in direct orientation, including a 30-kbp block containing *BOLA2* and two other genes and a 72-kbp block harbouring a partial segmental duplication of *SMG1* (*SMG1P*). Expansion and contraction of this cassette underlie hundreds of kilobase pairs of structural diversity between human haplotypes. *BOLA2* paralogue-specific copy number genotype data suggest that H1 and H3 probably represent the most common haplotype structures in humans.

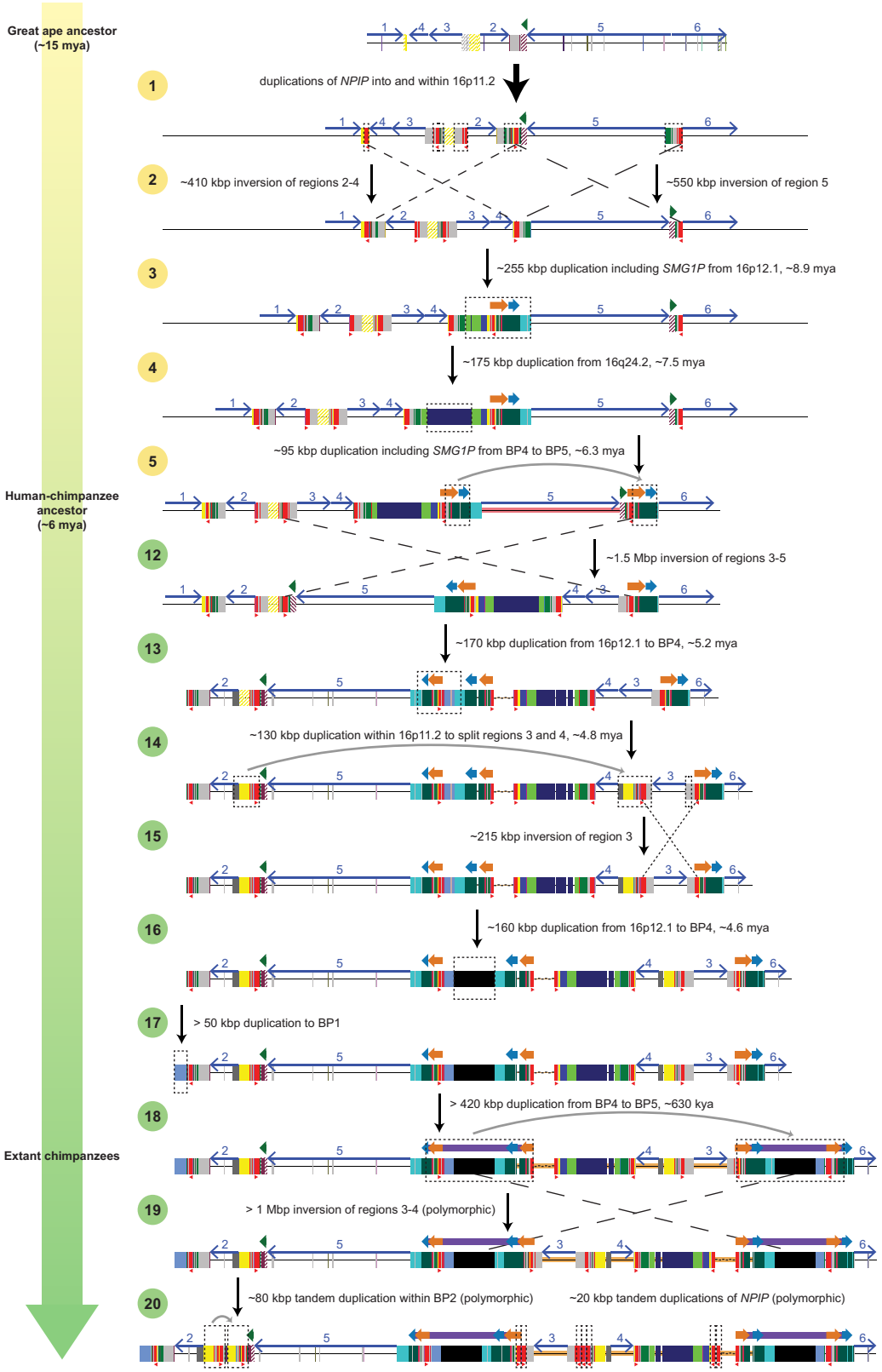


Extended Data Figure 2 | See next page for caption.

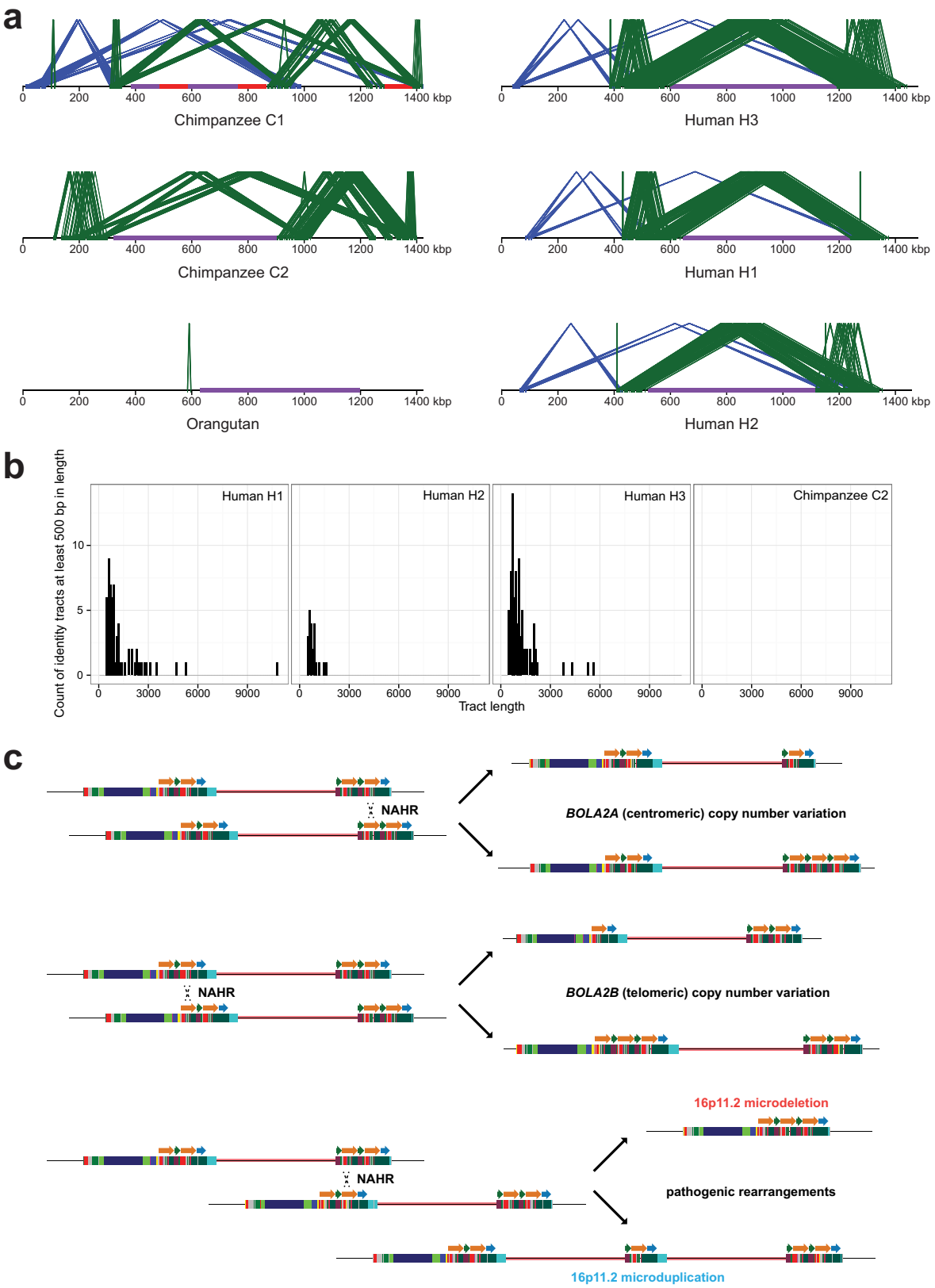
Extended Data Figure 2 | Comparison of chromosome 16p11.2 structure between apes. a, Sequences (thin horizontal lines) from human (GRCh37 chr16:28195661–30573128) and orangutan (contig sequence) at chromosome 16p11.2 are compared using Miropeats ($s = 1{,}000$) and annotated with locations of human segmental duplications and FISH probes used to validate the organization of the region. Lines connecting the sequences show regions of homology, and line colours highlight differences in the order and orientation of distinct gene-rich regions of unique sequence across the locus (numbered 1–6). Numbers below FISH probes correspond to numbers within the images on the right, specifying which probes were used in each experiment. Experiment 1 used the same probes as experiment 3, and experiment 2 used the same probes as experiment 4. Three-colour interphase FISH on human and orangutan chromosomes confirms the accuracy of our assembled orangutan contig. **b**, Sequences (thin horizontal lines) from human (GRCh37 chr16:28195661–30573128) and two chimpanzee structural haplotypes at chromosome 16p11.2 are compared using Miropeats ($s = 1{,}500$) and annotated with locations of human segmental duplications and FISH probes used to validate the organization of the region. Thick red horizontal lines indicate gaps in the chimpanzee contigs, and black boxes correspond to chimpanzee-specific segmental duplications (that is, sequences not duplicated in humans). Lines connecting the sequences show regions of homology, and line colours highlight differences in the order and orientation of distinct gene-rich regions of unique sequence across the locus (numbered 2–6). Numbers below FISH probes correspond to numbers within the images on the right, specifying which probes were used in each experiment. Grey rectangles show mapping locations of FISH probes in human. Three-colour interphase FISH on chimpanzee chromosomes confirms the accuracy of our assembled contigs.



Extended Data Figure 3 | Dynamic evolution of human chromosome 16p11.2. a, A model for the evolution of the chromosome 16p11.2 BP1–BP5 region[8] during great ape evolution. The schematic depicts structural changes over time leading to the present-day human architecture (see Supplementary Information for details). The orangutan structure (top) is largely devoid of segmental duplications and deemed to represent the ape ancestral organization because it is conserved with mouse. Subsequent steps were inferred on the basis of phylogenetic reconstruction, origins of the duplicated sequences and the most parsimonious path with respect to changes in gene order (inversions). (See Supplementary Information for a detailed discussion of all supporting evidence and confidence levels for each step.) Note that, without access to genomes containing intermediate chromosome 16p11.2 structures, it is impossible to know with certainty the entire step-by-step evolutionary history. Some details presented here may not be accurate. mya, million years ago; kya, thousand years ago.

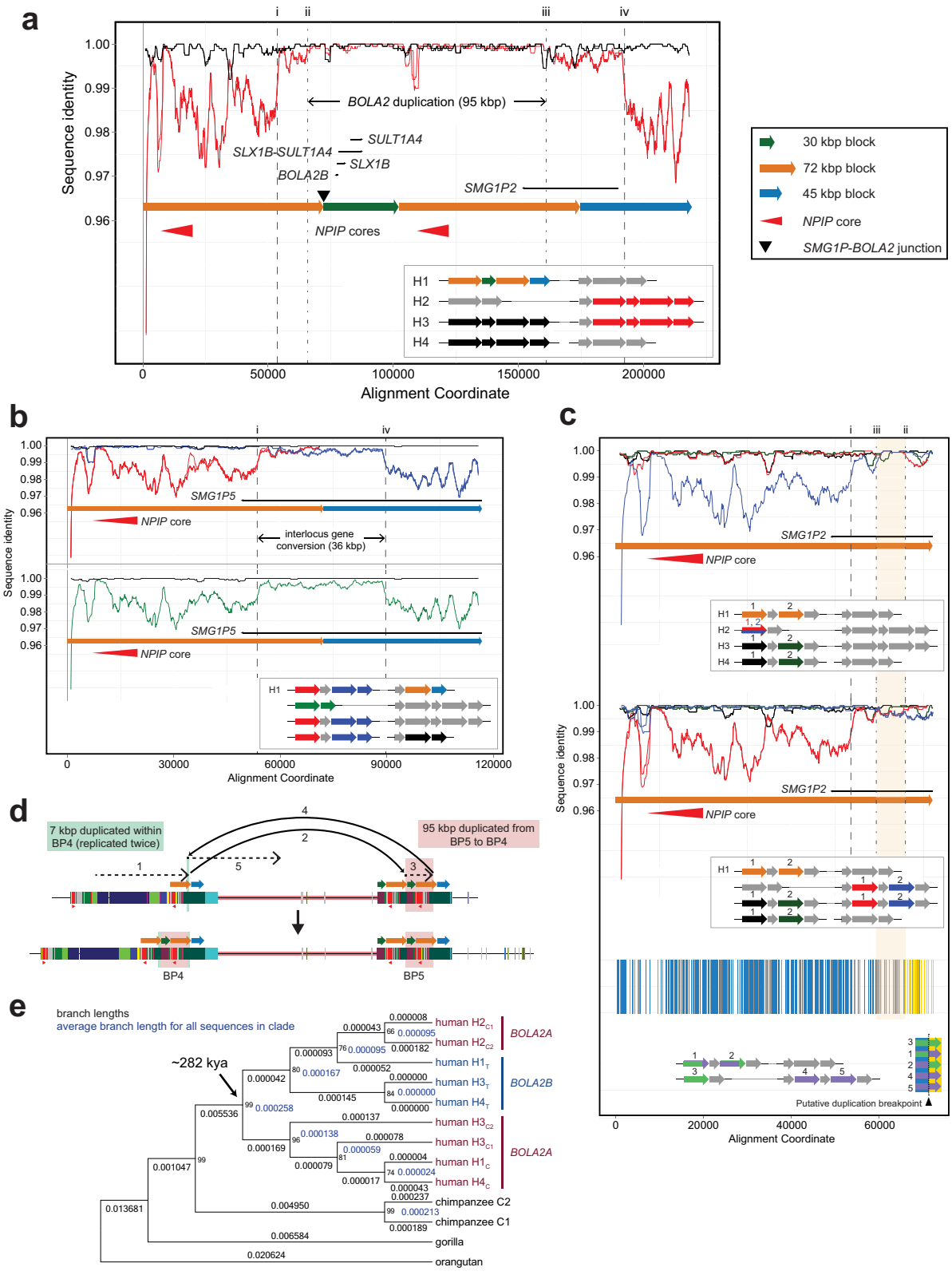


Extended Data Figure 4 | Dynamic evolution of chimpanzee chromosome 16p11.2. A model for the evolution of the chromosome 16p11.2 BP1–BP5 region[8] during great ape evolution. The schematic depicts structural changes over time leading to the present-day chimpanzee architecture (see Supplementary Information for details and discussion of all supporting evidence and confidence levels for each step).



Extended Data Figure 5 | See next page for caption.

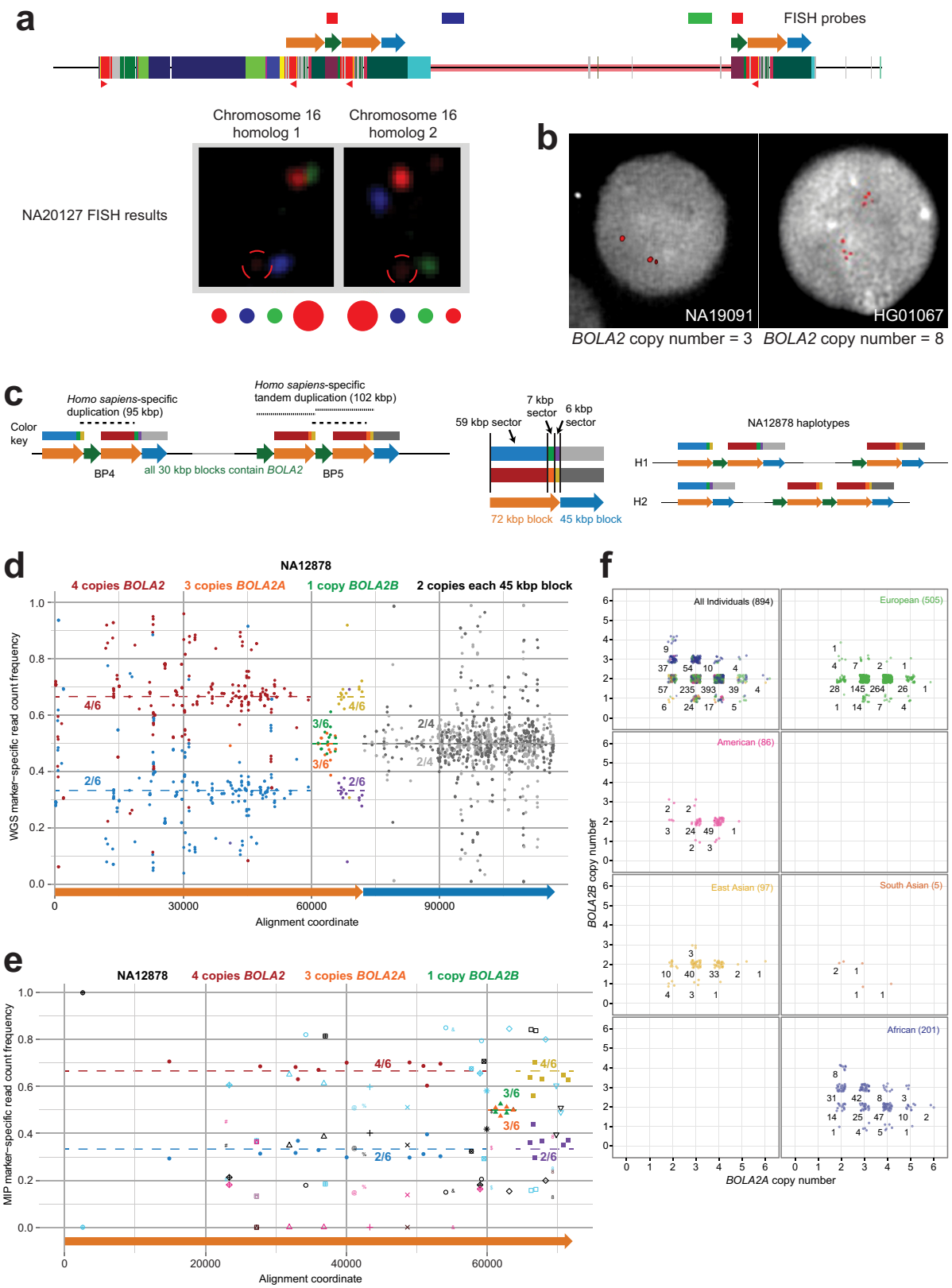
Extended Data Figure 5 | Comparison of duplications around the chromosome 16p11.2 autism critical region among apes and nonallelic homologous recombination (NAHR) model underlying CNV at human chromosome 16p11.2. a, Local directly oriented (green) and inversely oriented (blue) intrachromosomal segmental duplications flanking the chromosome 16p11.2 autism critical region (purple) are visualized using Miropeats (s = 1,000). Gaps in the chimpanzee C1 contig are shown in red. The smaller size (<50 kbp) and lower average sequence identity (at most 98.6%) of directly oriented duplications flanking the critical region in chimpanzee compared with human haplotypes including *BOLA2* on both sides of the critical region (at least 147 kbp of directly oriented duplications having at least 99.3% average sequence identity) suggest that susceptibility to NAHR resulting in microdeletions and microduplications at this locus evolved specifically in humans. **b**, Perfect sequence identity tract lengths (>500 bp) within directly oriented duplications flanking the critical region for human versus chimpanzee. Histograms show counts of tracts of perfect sequence identity (lacking single-nucleotide variants and indels) between directly oriented segmental duplications of interest within each indicated haplotype and the distribution of these tracts over different size ranges. Human haplotypes having *BOLA2* on both sides of the critical region (H1 and H3) contain the highest number of such tracts and the longest such tracts, including one tract spanning 10,774 bp. In contrast, the longest tract of perfect sequence identity between duplications of interest in chimpanzee (considering both the C1 and C2 haplotypes) spans 450 bp. **c**, NAHR model underlying normal and disease-associated CNV at human chromosome 16p11.2.



T = ((0.000258 subs/site)/(0.000258+0.005536+0.004950+0.000213 subs/site))*(12 million years) = ~0.282 million years

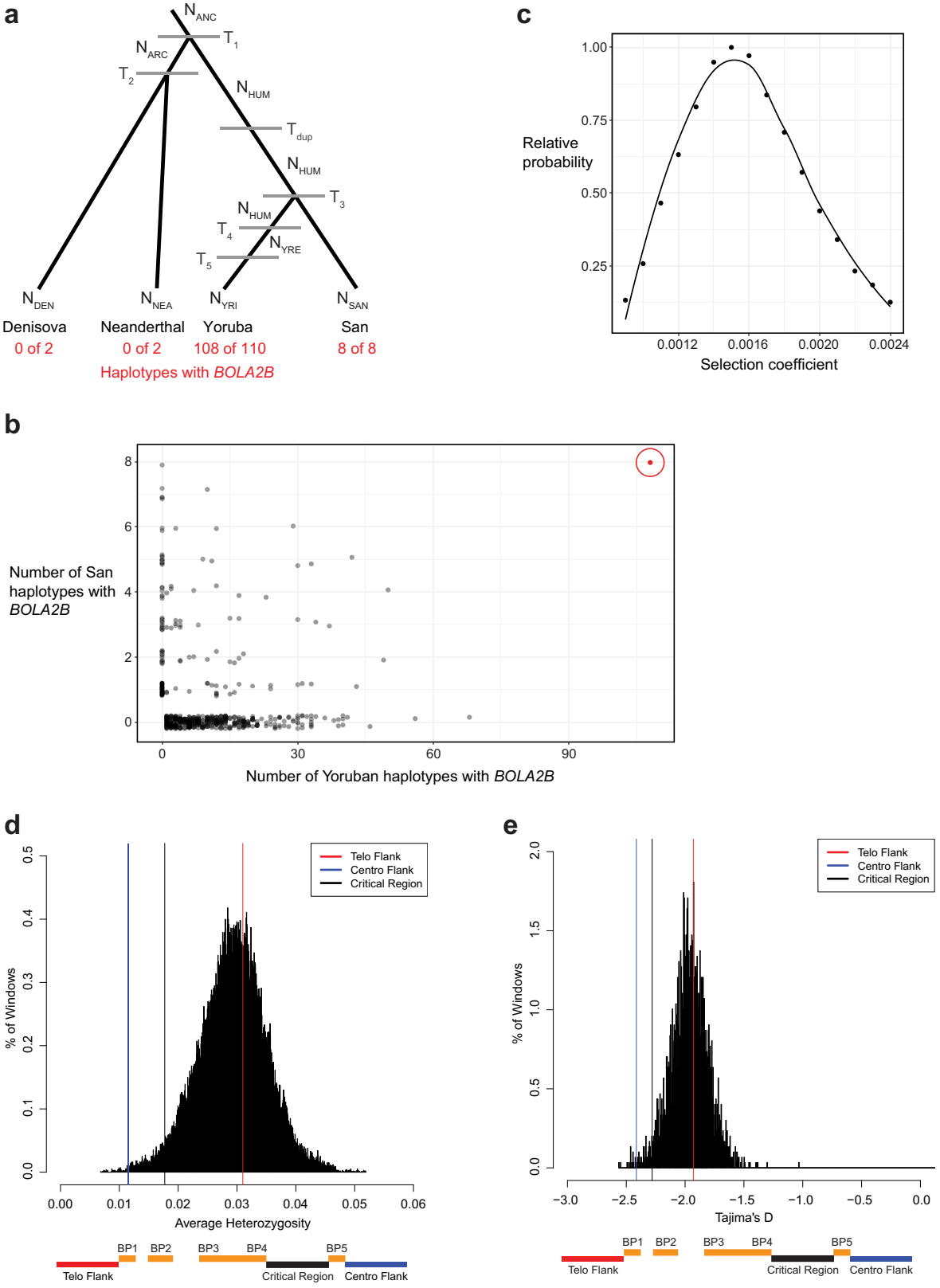
Extended Data Figure 6 | See next page for caption.

Extended Data Figure 6 | Sequence refinement of interspersed *BOLA2* duplication breakpoints, inference of *BOLA2* duplication mechanism and phylogenetic *BOLA2* duplication timing. a, H1 human BP4 sequence (orange, green, orange and blue arrows in inset) was aligned to its allelic (black arrows in inset) and paralogous (red arrows in inset) counterparts. The sequence identity for each alignment was computed and plotted over 2-kbp windows, sliding by 100 bp. Black lines indicate sequence identity for allelic comparisons, whereas red lines correspond to paralogous comparisons. While the allelic comparisons exhibit uniform, near-perfect sequence identity across the entirety of the alignment, paralogous comparisons reveal three distinct levels of sequence identity, with the highest level in the middle. This pattern suggests that the *BOLA2* duplication (highest-identity region, 95 kbp) landed within an evolutionarily older segmental duplication having paralogues at BP4 and BP5. Dashed vertical lines (numbered i–iv) indicate putative breakpoints for events that occurred after this older segmental duplication. Junction sequence from the BP5 102-kbp tandem duplication (that is, the *SMG1P*–*BOLA2* junction) was clearly included in the 95-kbp duplication from BP5 to BP4. **b**, Alignment of BP4 sequences containing the putative left (red arrows in inset) and right (dark blue arrows in inset) *BOLA2* duplication breakpoints to the BP5 paralogue associated with the evolutionarily older segmental duplication (orange and light blue arrows in inset) and sliding window sequence identity analysis supports the hypothesis outlined above. Sequence identity lines for comparisons involving left and right BP4 sequences intersect in the vicinity of the hypothesized *BOLA2* duplication breakpoints. Comparing this result with the same analysis of the human H2 BP4 sequence lacking *BOLA2* (green arrows in inset and green identity line) suggests this BP4 sequence represents the ancestral state of BP4 before the *BOLA2* duplication arrived. Thus, two levels of sequence identity existed between BP4 and BP5 before the *BOLA2* duplication, consistent with an interlocus gene conversion event. **c**, Alignment of BP4 sequences (orange arrows in insets) containing the putative *BOLA2* duplication breakpoints to their ancestral BP4 (top plot) and their ancestral BP5 (middle plot) counterparts and sliding window sequence identity analysis reveals an ~7-kbp window (highlighted in orange) defining the *BOLA2* duplication breakpoints. Analysis of the underlying multiple sequence alignment (Supplementary Table 5) identified positions with signatures informative for breakpoint localization (blue vertical lines, left BP4 72-kbp block outside the *BOLA2* duplication and right BP4 72-kbp block within the *BOLA2* duplication; yellow vertical lines, left BP4 72-kbp block within the *BOLA2* duplication and right BP4 72-kbp block outside the *BOLA2* duplication). Grey vertical lines indicate positions showing signatures of interlocus gene conversion. As both left and right 72-kbp block BP4 sequences within the ~7-kbp window are more highly identical to ancestral BP4 sequence (20/24 informative positions match the ancestral BP4 sequence) than to ancestral BP5 sequence, it is likely that this interval was involved in the *BOLA2* duplication but duplicated only within BP4. Its boundaries define the most likely *BOLA2* duplication breakpoints, and this pattern of sequence identity suggests a template-switching replicative mechanism as most probably underlying the *BOLA2* duplication event. **d**, Template-switching model for the formation of *BOLA2B*. This mechanism was inferred from the sequence identity analyses in **a**–**c** and from analysis of a multiple sequence alignment (Supplementary Table 5). **e**, Phylogenetic characterization of the 95-kbp duplication containing *BOLA2* from BP5 to BP4. Cladogram representation of an unrooted neighbour-joining phylogenetic tree based on a 21,102-bp multiple sequence alignment spanning *BOLA2* and most of the 30-kbp block including human sequences from BP4 and BP5 and single-copy orthologous sequences from chimpanzee, gorilla and orangutan. Branch lengths (substitutions per site) are shown on each branch (black decimal numbers), and bootstrap support is indicated (black integers at nodes). Blue numbers correspond to nodes and indicate average branch lengths for all sequences in corresponding clades. Branch lengths were used to estimate the time corresponding to the 95-kbp duplication containing *BOLA2* from BP5 to BP4 as shown.



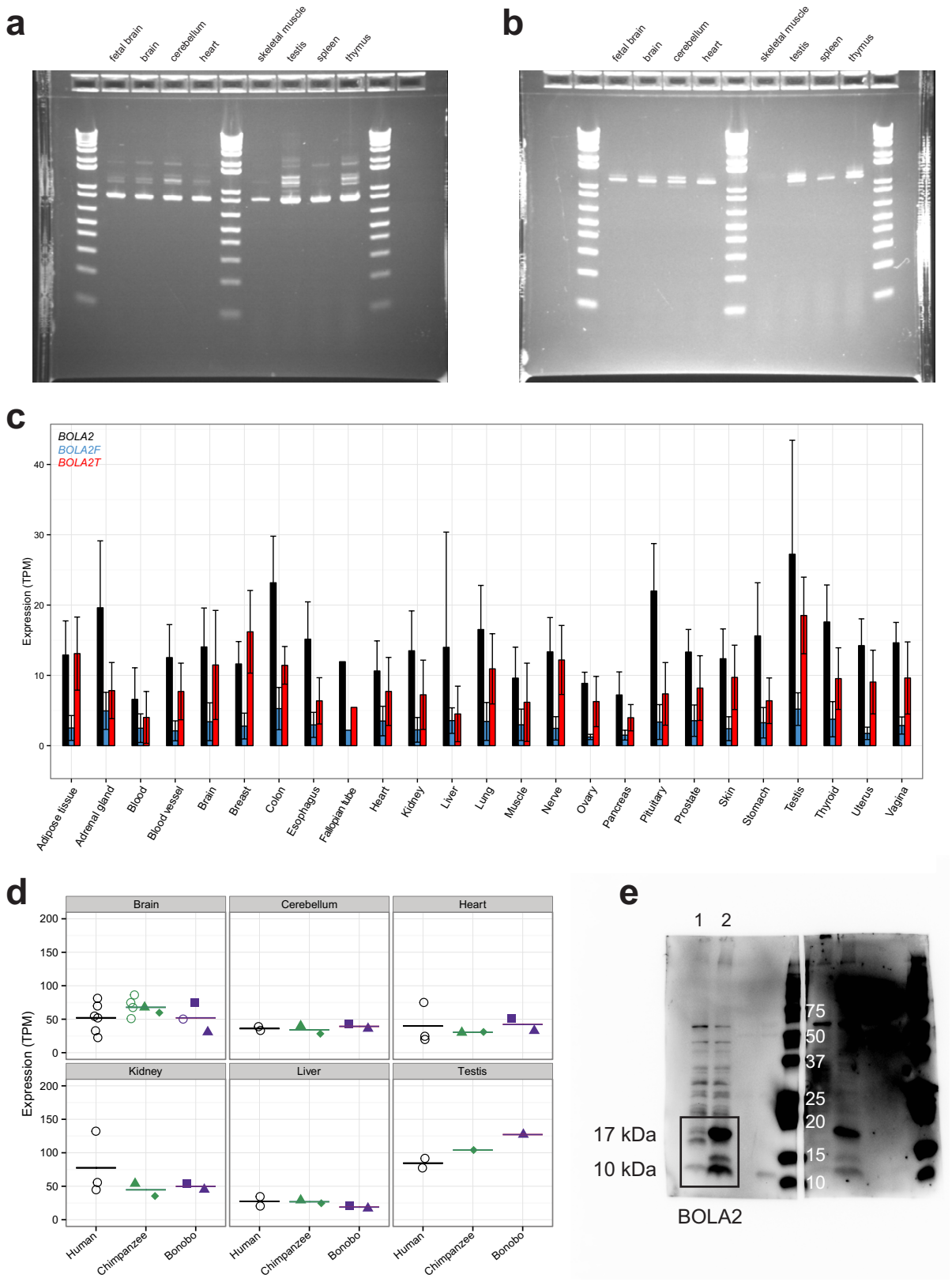
Extended Data Figure 7 | See next page for caption.

Extended Data Figure 7 | Analyses of *BOLA2* aggregate and paralogue-specific CNV in humans. a, Interphase FISH confirms both *BOLA2A* and *BOLA2B* show CNV. Previous interphase FISH analysis (data not shown) suggests the individual NA20127 has six total copies of *BOLA2*. Diagram outlines a three-colour FISH assay including two probes (blue, green) targeting sequences within the autism critical region and one probe (red) targeting ~18-kbp of sequence (including *BOLA2*) over the 30-kbp duplication block. Signals from the red probe are detected on the telomeric (BP4) and centromeric (BP5) sides of the critical region (adjacent to the blue and green probes, respectively) on both chromosome 16 homologues. However, the red probe signal intensity is strongest adjacent to the green probe for one homologue but, in contrast, is strongest adjacent to the blue probe for the other chromosome 16 homologue, consistent with higher *BOLA2A* copy number in the first case and higher *BOLA2B* copy number in the second case. These data indicate that individual NA20127 has three copies each of *BOLA2A* and *BOLA2B*. This differential signal intensity pattern does not result from an inversion of the chromosome 16p11.2 critical region in this individual, as data from another FISH experiment (data not shown) refute this possibility. Information on probes used in these FISH experiments is provided in Supplementary Table 2. **b**, Interphase FISH experiments using a probe targeting *BOLA2* and surrounding sequence for individuals having the lowest (three) and highest (eight) confirmed aggregate *BOLA2* copy numbers. **c**, Left and middle schematics detail three distinct sectors of the 72-kbp blocks (orange arrows). Each block has paralogous sequence variants that are informative for particular region(s) compared with others in chromosome 16p11.2. These markers are colour-coded into three sectors within the 72-kbp block of paralogy (a 59-kbp sector, blue and red boxes; a 7-kbp sector, green and orange boxes; and a 6-kbp sector, purple and yellow boxes), indicating which particular regions they distinguish. Right schematic shows known haplotype structures for individual NA12878. **d**, Analysing WGS data from NA12878 yields copy number estimates for *BOLA2A* and *BOLA2B* that match the known *BOLA2* paralogue-specific copy number (PSCN) for this individual. Each point shows a relative marker-specific read count frequency (*y* axis) and its position within the duplication blocks (*x* axis). Point colours correspond to different marker sets for each sector, as diagrammed in **c**. Fractions indicate the relative copy number of each marker set. Estimates of 4/6 (red marker set) versus 2/6 (blue marker set) for the 59-kbp sector confirms the sequenced architecture (**c**) with an aggregate of four *BOLA2* copies, and the estimate of 3/6 (orange marker set) confirms three copies of *BOLA2A*. WGS analysis also yields accurate PSCN estimates for the 45-kbp block. **e**, Using MIPs, we employed the same relative read-depth strategy. Genotyping results for the same sample as in **d** are shown, with additional markers (points not colour-coded as in **c** and **d**) added on the basis of polymorphic variants (symbols indicate different patterns of presence/absence among 72-kbp blocks, considering all such blocks from our four contiguous human haplotypes). MIP genotypes confirm WGS estimates (in **d**). **f**, *BOLA2* PSCN genotypes (points, jittered around their integer values for clarity) were inferred from MIP sequence data for 894 humans. Numbers indicate total counts of individuals in each population having a particular *BOLA2* PSCN genotype. Low-confidence estimates were excluded.



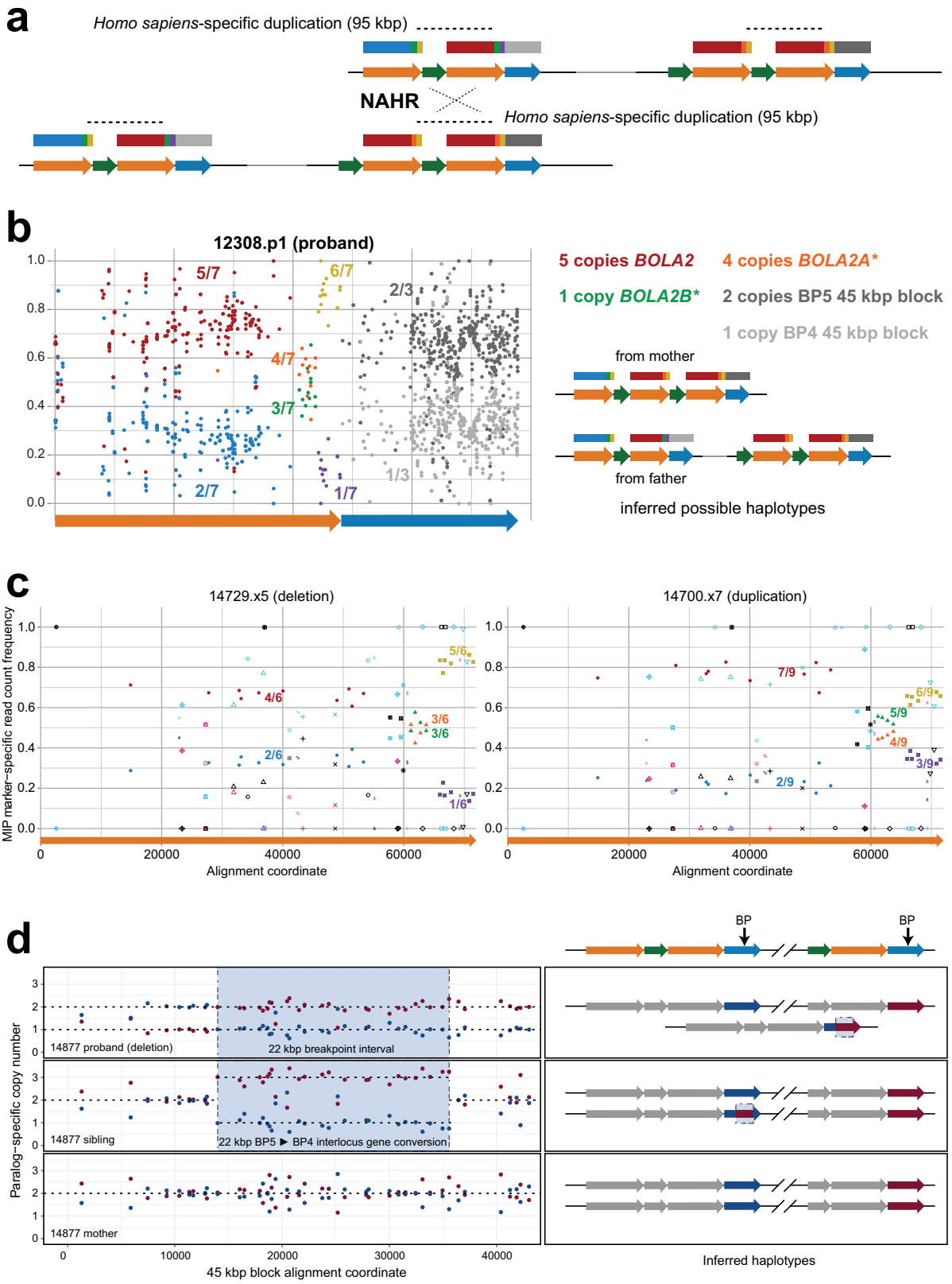
Extended Data Figure 8 | See next page for caption.

Extended Data Figure 8 | Population genetic modelling of the *BOLA2B* duplication and critical region analyses. a, Demographic model (adapted from ref. 16) used to simulate *BOLA2B* evolution under different scenarios. N_{ANC}, effective population size of *Homo* ancestor, 21,600. N_{ARC}, effective population size of Neanderthal-Denisova ancestor, 500. N_{HUM}, effective population size of human ancestor, 24,000. N_{YRE}, effective size of Yoruban population after expansion, 45,000. N_{DEN}, effective population size of Denisova, 500. N_{NEA}, effective population size of Neanderthal, 500. N_{YRI}, effective size of extant Yoruban population, 10,000. N_{SAN}, effective size of extant San population, 10,000. T_1, time of archaic hominin divergence from modern humans, 650,000 years. T_2, time of Neanderthal–Denisova divergence, 525,000 years. T_{dup}, time of formation of *BOLA2B*, 282,000 years. T_3, time of Yoruban-San divergence, 200,000 years. T_4, time of Yoruban population expansion, 157,500 years. T_5, time of Yoruban population decline, 37,500 years. **b**, Simulation results ($n = 1{,}000{,}000$) assuming that the duplication that formed *BOLA2B* occurred once, 282 ka, along the modern human ancestral lineage and evolved under neutrality compared with the observed genotype frequencies of *BOLA2B* in 8 San and 110 Yoruban haplotypes. Nearly all (999,531) simulations resulted in *BOLA2B* being lost from both populations; results from the remaining 469 simulations (black) are shown alongside the observed data (red, circled). Under this simple neutral model incorporating *BOLA2B* age, the observed *BOLA2B* frequency is never approached. **c**, Simulation was repeated exploring a range of selection coefficients from 0.0009 to 0.0024 (increments of 0.0001), and the relative probability of the observed data under each scenario was calculated as the proportion of simulations yielding the observed *BOLA2B* genotypes among simulations where *BOLA2B* was not lost relative to the maximum such proportion for any single selection coefficient considered. The maximum likelihood estimate for the selection coefficient was $s = 0.0015$. Smoothed line is the LOESS regression curve. **d**, Low average heterozygosity of the chromosome 16p11.2 BP4–BP5 critical region. Distribution of average heterozygosity values for 100,000 ~550-kbp regions of unique sequence randomly sampled with replacement from the autosomal genome compared with average heterozygosity values for the critical region (black line) and flanking unique sequences (coloured lines). The critical region lies in the bottom 2.6% of the distribution, showing low diversity consistent with potential positive selection. Bottom schematic indicates locations of the critical region and flanking unique regions in relation to segmental duplications across the locus—note that *BOLA2A* is located at BP5 and *BOLA2B* at BP4. Telo, telomeric; Centro, centromeric. **e**, Low Tajima's *D* score for the chromosome 16p11.2 BP4–BP5 critical region. Distribution of Tajima's *D* scores for 2,987 non-overlapping ~550-kbp regions across the genome compared with Tajima's *D* scores for the critical region (black line) and flanking unique sequences (coloured lines). The critical region lies in the bottom 2.7% of the distribution, consistent with possible positive selection. The distribution is centred near −2 rather than 0 because most single-nucleotide variants in the 1000 Genomes Project data set are rare variants having arisen during the large expansions of human populations over the past 100,000 years. Bottom schematic indicates locations of the critical region and flanking unique regions in relation to segmental duplications across the locus.



Extended Data Figure 9 | See next page for caption.

Extended Data Figure 9 | *BOLA2* expression and antibody validation. **a**, RT–PCR expression profile for canonical *BOLA2*. The expected product size for canonical *BOLA2* (838 bp) was observed in all eight human tissues; 1 kb + DNA ladder (Thermo Fisher). **b**, RT–PCR expression profile for *BOLA2-SMG1* fusion product. The expected product size for the *BOLA2* fusion transcript (1,239 bp) was observed as a doublet in all tissues except skeletal muscle. Intensity of upper band differs between tissues; 1 kb + DNA ladder (Thermo Fisher). **c**, *BOLA2* RNA-seq expression analysis. Canonical (*BOLA2*) and fusion transcripts (*BOLA2F*, *BOLA2T*) were assessed across 25 humans from GTEx RNA-seq data. Bar heights indicate mean expression levels for each tissue in transcripts per million with standard errors shown (error bars). Colours correspond to different *BOLA2* isoforms as indicated. **d**, *BOLA2* expression among primates in six adult tissues. Each point indicates a *BOLA2* expression estimate from a single tissue sample, with samples obtained from a total of 18 humans, 6 chimpanzees and 3 bonobos. Open circles correspond to individuals analysed in a single experiment, while closed shapes denote data from multiple experiments involving the same individual, with each distinct colour plus shape pattern showing all experiments for a particular individual. Horizontal lines show mean expression values for each species and tissue. Combined with our expression analyses of iPSCs, these data show *BOLA2* expression differs substantially between human, chimpanzee and bonobo only in stem cells. **e**, Western blotting of HeLa cells transfected with the human *BOLA2* annotated CDS and probed with an anti-BOLA2 antibody (Sc-163747). Whole-cell lysates of HeLa cells non-transfected with the overexpression construct (lane 1) and transfected with the human *BOLA2* annotated CDS (lane 2) were probed with anti-BOLA2 antibody. Two bands with molecular weights of 10 and 17 kDa are identified, are more abundant in transfected cells and correspond to two BOLA2 protein isoforms arising from different translation start sites.



Extended Data Figure 10 | See next page for caption.

Extended Data Figure 10 | Chromosome 16p11.2 rearrangement breakpoint refinement. a, NAHR between directly oriented segmental duplications at BP4 and BP5. This unequal crossover results in chromosome 16p11.2 microdeletions and microduplications (Extended Data Fig. 5c). Coloured arrows and boxes correspond to duplication blocks and sectors within them are colour-coded as in Extended Data Fig. 7c. Unequal crossover could occur in eight distinct regions with regard to duplication block and sector boundaries. Three such regions are located within the ~95-kbp *H. sapiens*-specific duplication (dashed lines). Only unequal crossover events outside the *H. sapiens*-specific duplication produce recombinants which have a sector with non-uniform marker-specific copy number across its extent. **b**, Relative marker-specific read count frequencies (points) determined from WGS analysis for a microdeletion proband. Fractions indicate relative marker-specific copy numbers, as in Extended Data Fig. 7d, and diagrams adjacent to the plot show inferred haplotype structures for each chromosome 16 homologue for this individual. Although the data in the plot provide only diploid genotypes (and not resolved haplotypes), the haplotypes suggested here reflect this genotype information together with data from the parents (not shown) and the assumption (supported by our PSCN data) that haplotypes which have two *BOLA2A* copies and a single *BOLA2B* copy are the most common. Because marker-specific copy number is uniform across each sector, unequal crossover breakpoints must have occurred within the *H. sapiens*-specific duplication. **c**, Breakpoint refinement based on MIP PSCN marker data. Plots show relative marker-specific read count frequencies (points) determined using MIPs for a typical microdeletion patient (left) and a typical microduplication patient (right). Shapes and colour code designate different markers, and fractions indicate relative marker-specific copy numbers (as in Extended Data Fig. 7). Because marker-specific copy number is uniform across each sector for both individuals, in both cases, unequal crossover breakpoints must have occurred within the *H. sapiens*-specific duplication. **d**, Data from an atypical patient where the breakpoints are inferred to map outside the *H. sapiens*-specific segmental duplication. The plots show paralogue-specific copy number for a chromosome 16p11.2 microdeletion proband, his sibling and his mother over a 45-kbp duplication block shared between BP4 and BP5. Paralogue-specific copy number was estimated using a MIP assay targeting 54 informative markers over this region, with data from 43 markers fixed among haplotypes H1–H4 shown (points). Dashed lines indicate calls inferred using an automated caller, which were also confirmed by visual inspection. Adjacent schematics indicate the inferred haplotypes for each individual on the basis of these data, with approximate breakpoint locations shown (arrows). The results demarcate the location of the unequal crossover interval on the basis of the reciprocal copy number transition between the BP5 (red) and BP4 (blue) 45-kbp block segmental duplications. In this case, the breakpoints clearly map to a 22-kbp region outside the typical hotspot. Analysis of the sibling suggests that this region was the site of an interlocus gene conversion event from BP5 to BP4, and data from the mother imply that chromosomes having this event were present in the paternal germline. DNA from the father was not available for testing.

LETTER

doi:10.1038/nature19058

A trans-synaptic nanocolumn aligns neurotransmitter release to receptors

Ai-Hui Tang[1,2]*, Haiwen Chen[1,2,3]*, Tuo P. Li[1,2,3], Sarah R. Metzbower[1,2], Harold D. MacGillavry[4] & Thomas A. Blanpied[1,2]

Synaptic transmission is maintained by a delicate, sub-synaptic molecular architecture, and even mild alterations in synapse structure drive functional changes during experience-dependent plasticity and pathological disorders[1,2]. Key to this architecture is how the distribution of presynaptic vesicle fusion sites corresponds to the position of receptors in the postsynaptic density. However, while it has long been recognized that this spatial relationship modulates synaptic strength[3], it has not been precisely described, owing in part to the limited resolution of light microscopy. Using localization microscopy, here we show that key proteins mediating vesicle priming and fusion are mutually co-enriched within nanometre-scale subregions of the presynaptic active zone. Through development of a new method to map vesicle fusion positions within single synapses in cultured rat hippocampal neurons, we find that action-potential-evoked fusion is guided by this protein gradient and occurs preferentially in confined areas with higher local density of Rab3-interacting molecule (RIM) within the active zones. These presynaptic RIM nanoclusters closely align with concentrated postsynaptic receptors and scaffolding proteins[4–6], suggesting the existence of a trans-synaptic molecular 'nanocolumn'. Thus, we propose that the nanoarchitecture of the active zone directs action-potential-evoked vesicle fusion to occur preferentially at sites directly opposing postsynaptic receptor–scaffold ensembles. Remarkably, NMDA receptor activation triggered distinct phases of plasticity in which postsynaptic reorganization was followed by trans-synaptic nanoscale realignment. This architecture suggests a simple organizational principle of central nervous system synapses to maintain and modulate synaptic efficiency.

The location of vesicle fusion within an active zone is probably dictated by a few key members of the presynaptic proteome, including RIM1/2, Munc13, and bassoon (Bsn)[7] (Fig. 1a). To explore the organization of these proteins, we studied their subsynaptic distribution relative to postsynaptic scaffolding protein PSD-95 in cultured hippocampal neurons using 3D-STORM[8] following immunolabelling using primary antibodies and Alexa647- or Cy3-tagged secondary antibodies (Fig. 1b). Paired synaptic clusters of active zone protein and PSD-95 with clear borders were selected. As a confirmation that these pairs constituted synapses, we measured the peak-to-peak distances between pre- and postsynaptic clusters and found them to be consistent with previous measurements[9] (Extended Data Fig. 1).

The distribution of RIM1/2 within the active zone, measured as 3D local density, was distinctively non-uniform with notable high-density peaks, which we characterized as nanoclusters (Fig. 1c, e). We adapted an auto-correlation function (ACF) to test whether this distribution occurs more frequently than expected by chance. The measured ACF showed significant non-uniformity compared to random ensembles (Fig. 1d). Simulations showed that the distance for which the ACF was significantly elevated provided a means to estimate the nanocluster diameter (Extended Data Fig. 2a–c). The average estimated diameter of ~80 nm for RIM1/2 nanoclusters was very close to the reported size of PSD-95 and AMPA receptor (AMPAR) nanoclusters[4–6]. Similar distribution and nanocluster properties were found using a different antibody targeted towards a separate epitope in RIM1 (Extended Data Fig. 2d). Isolated non-synaptic small groups of localizations showed a weaker ACF that was significant over a much smaller distance (Fig. 1d). This and other experiments suggest that the measured non-uniformity was not likely due to over-counting molecules or to potential artefacts of primary–secondary antibody labelling (Extended Data Fig. 3).

To directly compare the nanoscale organization of key active zone proteins, we developed an algorithm that identified nanoclusters based on local densities (Fig. 1e). Nanoclusters of each protein were more likely to be located near the centre of synapses than near the edge (Fig. 1f, Extended Data Fig. 2i). Compared to PSD-95 as the common control in pairwise two-colour experiments, there were similar numbers of RIM1/2, more Munc13, and fewer Bsn nanoclusters per synapse (Fig. 1h). Comparisons between these three proteins suggested that Munc13 had a wider distribution than RIM1/2 across the active zone and the distribution of Bsn was closer to uniform throughout the synapse (Fig. 1g–i, Extended Data Fig. 2f–n). Together, these observations revealed a complex and heterogeneous molecular architecture within single synapses, typified by dense assemblies of fusion-associated proteins nearer the centre.

To examine the potential functional impact of the active zone nanoclusters on vesicle fusion[10,11], we sought to directly map the distribution of vesicle fusion sites over multiple release events within individual boutons. To do so, we adapted analysis for single-molecule localization to signals from single-vesicle fusion obtained with vGlut1–pHluorin–mCherry (vGpH). Neurons were cotransfected with cyan fluorescent protein (CFP)-tagged synapsin1a (Syn1a), a vesicle-associated protein that marks boutons, and vGpH, which increases in green fluorescence intensity upon vesicle fusion[12]. Single electrical field stimuli evoked vesicle fusion (Fig. 2a, b, Extended Data Fig. 4a) with a release probability (P_r) of 0.11 ± 0.01 (mean $\pm$ s.e.m.) per bouton, comparable to previous measurements, which was also sensitive to extracellular Ca^{2+} (Extended Data Fig. 4b–d), as expected. In the presence of TTX, the frequency of action-potential-independent spontaneous release events detected with vGpH was similar to the frequency of NMDA receptor (NMDAR)-dependent postsynaptic Ca^{2+} transients measured separately using the Ca^{2+} sensor GCaMP6f (Extended Data Fig. 5a).

To determine whether these evoked fusion events represent single- or multi-vesicular fusion, we compared them with spontaneous release under TTX conditions (Fig. 2a–c), which most likely arises from single vesicle fusion[13]. By fitting the photon number distributions of evoked and spontaneous events, we estimated that ~72–82% of evoked events arose from single-vesicle fusion (Fig. 2c). With the majority of evoked release stemming from single-vesicle fusion, the location of fusion may be deduced by mathematically fitting the fluorescence profile captured immediately after fusion (Fig. 2d), analogous to single-molecule

[1]Department of Physiology, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA. [2]Program in Neuroscience, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA. [3]Medical Scientist Training Program, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA. [4]Cell Biology, Department of Biology, Faculty of Science, Utrecht University, 3584 CH Utrecht, The Netherlands.
*These authors contributed equally to this work.

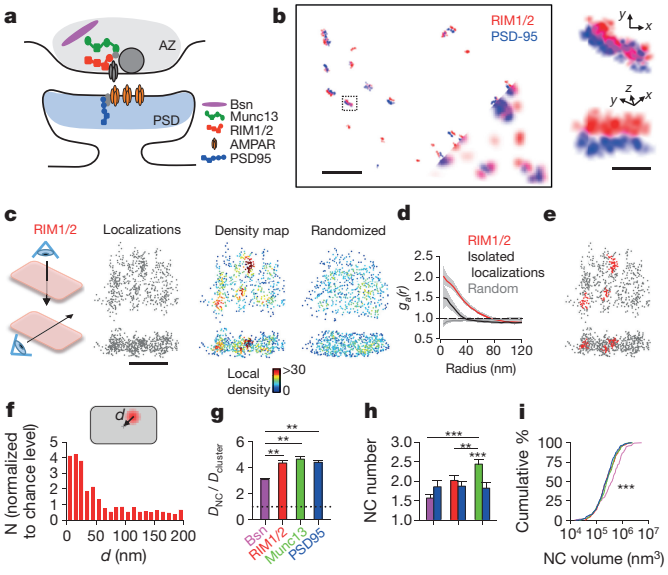


Figure 1 | Vesicle release proteins form subsynaptic nanoclusters. **a**, Colour-coded schematic of studied synaptic proteins. AZ, active zone; PSD, postsynaptic density. **b**, Synapses labelled with RIM1/2 and PSD-95 imaged using 3D-STORM (10-nm pixels) compared to wide-field composite (bottom corner, 100-nm pixels). Scale bar, 2 μm. Boxed synapse enlarged in original (top) and rotated (bottom) angles. Scale bar, 200 nm. **c**, *En face* (top) and side (bottom) views of a RIM1/2 cluster showing all localizations and local density maps for a measured synaptic cluster compared to a simulated randomized cluster. Scale bar, 200 nm. **d**, Auto-correlation functions of measured RIM1/2 ($n = 115$), isolated non-synaptic small groups of localizations due to repetitive switching of fluorophores ($n = 42$), and simulated randomized ($n = 115$) distributions. **e**, RIM1/2 nanoclusters (red) within a synaptic cluster. **f**, Distribution of nanocluster distances from the centre of synapses normalized to randomized distribution. **g**, Molecule density inside nanoclusters (NC) normalized to synaptic average. **h**, Average number of protein nanoclusters per synapse. **i**, Cumulative distributions of nanocluster volumes. $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$, one-way ANOVA on ranks with pairwise comparison procedures (Dunn's method) for **g**, **h** and Kolmogorov–Smirnov test for **i**. All experiments were repeated ≥3 times. Also see Extended Data Fig. 3 and Supplementary Table 1.

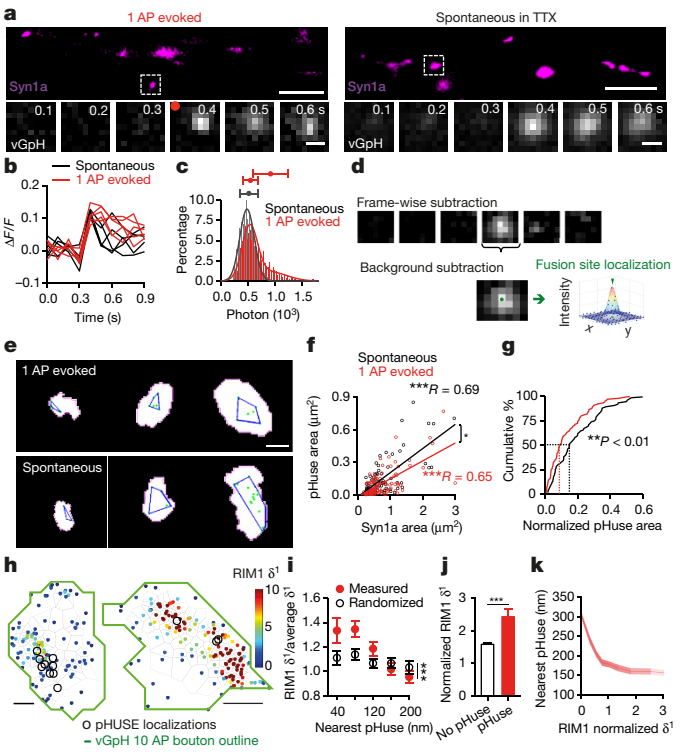


Figure 2 | Release site mapping by pHuse in single synapses shows RIM predicts evoked fusion distribution. **a**, Neurons co-expressing Syn1a–CFP (top; scale bars, 5 μm), identifying synaptic boutons, and vGpH (bottom; scale bars, 500 nm), used to detect vesicle fusion with fluorescence increases from single action potential (AP)-evoked and spontaneous release. **b**, Example of fluorescence traces from evoked and spontaneous events over repeated trials at single boutons. **c**, Photon count distributions for detected spontaneous events fit with a normal distribution ($\mu = 512$, $\sigma = 167$) and evoked events fit with a mixture of two normal distributions ($\mu_1 = 542$, $\sigma_1 = 143$, $\mu_2 = 912$, $\sigma_2 = 319$). Filled circles with error bars show mean ± s.d. of normal curves. **d**, Image processing steps in pHuse to determine fusion site locations. **e**, Fusion sites (green points) and area of fusion (blue line) from boutons of different sizes defined with Syn1a (white). Scale bar, 500 nm. **f**, Correlation between fusion area and bouton size, linear fit. Correlations are significantly different, ANCOVA, $F_{1,171} = 5.01$. **g**, Cumulative distributions of fusion areas normalized to bouton size (Kolmogorov–Smirnov test, $^{**}D = 0.26$). **f**, **g**, $n_{spontaneous} = 77/22$, $n_{evoked} = 104/28$. **h**, Tessellated RIM1-mEos and pHuse localizations over the same boutons. Scale bars, 200 nm. **i**, Tesseler first-rank density (δ^1) for RIM1 measured versus randomized distributions as a function of distance from pHuse localizations. **j**, Comparison within boutons of average δ^1 for RIM1 localizations within 40 nm to a pHuse localization versus not. **k**, Average nearest pHuse distance as a function of RIM1 δ^1. **i**, **j**, $n = 26/13$ $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. *n* given in synapses/experiments. Also see Extended Data Figs 4–6.

localization techniques[14]. For our median count of 518 photons per localization, the effective localization precision was in practice limited by vesicle diameter. In individual boutons, multiple evoked or spontaneous single-vesicle fusion events were used to generate maps that defined the areas over which vesicle fusion occurred (Fig. 2e, Extended Data Fig. 4e–l). We called this approach 'pHluorin uncovering sites of exocytosis' or pHuse.

Fusion site areas for spontaneous and evoked vesicle fusion tightly correlated with bouton areas measured by Syn1a (Fig. 2f), as expected. However, the slopes of the correlations differed, even though the bouton sizes were similar between groups (Extended Data Fig. 5b). In fact, evoked fusion site areas were significantly smaller (median smaller by 48%) and occurred over a significantly smaller proportion of the bouton (median smaller by 39%) than spontaneous fusion (Fig. 2g, Extended Data Fig. 5c, d, h–j).

One interpretation is that the concentration of vesicle priming proteins in nanoclusters favours evoked fusion in these subregions of the active zone. This predicts that pHuse events would be associated with higher local RIM1 density and conversely that high local density of RIM1 increases the probability of nearby fusion. To assess these predictions, we mapped vesicle fusion sites relative to Eos3-tagged RIM1 using sequential PALM-pHuse imaging on the same live boutons (Fig. 2h, Extended Data Fig. 6d, e). As a local density metric for RIM1, we applied Voronoi tessellation and measured the first-rank density (δ^1) for each RIM1–mEos3 localization (as described in ref. 15).

The distribution of RIM1–mEos3 was non-uniform and contained nanoclusters with an average diameter of 80.95 ± 5.34 nm and 78.93 ± 5.85 nm using either an adapted SR-Tesseler analysis[15] or nearest neighbour distance analysis[4], respectively (Extended Data Fig. 6f), consistent with our 3D-STORM results (Fig. 1). We then compared δ^1 as a function of distance from the nearest pHuse localization for the measured RIM1 distributions versus randomized RIM1 distributions generated from the same number of localizations over the same area. Indeed, near pHuse sites, the average RIM1 δ^1 was significantly greater than chance (Fig. 2i). Furthermore, within individual boutons, RIM1 molecules within 40 nm of a pHuse location had significantly higher δ^1 than those further away (Fig. 2j). Conversely, considering all individual RIM1 localizations, the distance from the nearest pHuse localization decreased as a function of RIM1 δ^1 (Fig. 2k).

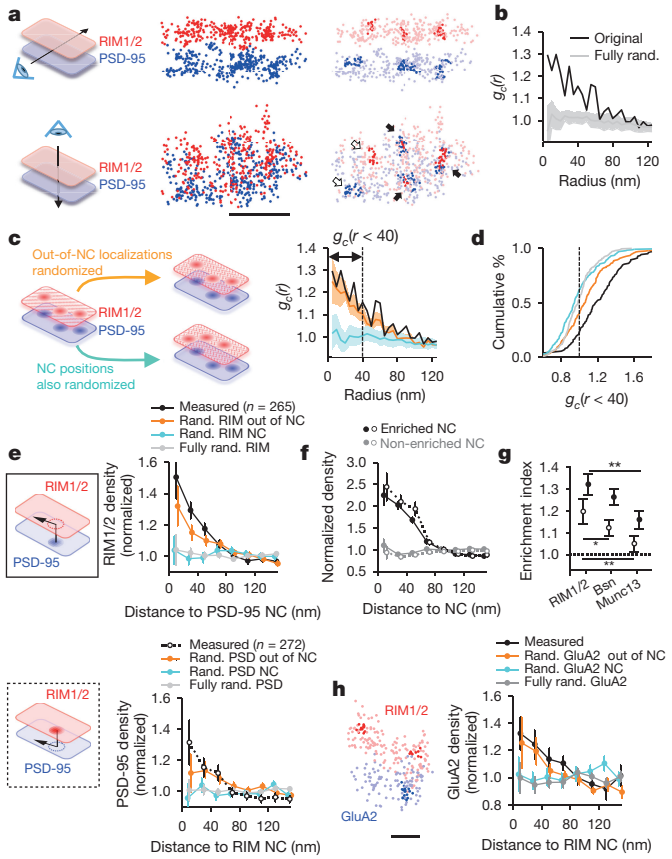


Figure 3 | Trans-synaptic nanoscale alignment of active zone and PSD proteins. a, Distributions of synaptic RIM1/2 and PSD-95 pair as the original localizations (left) and with nanoclusters highlighted (right). Scale bar, 200 nm. Filled arrows indicate aligned nanoclusters, open arrows denote non-aligned nanoclusters. **b**, Paired correlation function (PCF) of measured RIM1/2 and PSD-95 compared to PCF with either distribution randomized. **c**, PCF of simulated distributions with (cyan) and without (orange) shuffling nanocluster positions. **d**, Cumulative distributions of cross-correlation index ($n = 143$ synapses). **e**, RIM1/2 protein enrichment as a function of distance from translated PSD-95 nanocluster centres (top, filled points) and PSD-95 enrichment relative to RIM1/2 nanoclusters (bottom, open points). Simulations with same randomizations as in **d**, **e** were performed for each synapse. **f**, Protein density profile for enriched versus non-enriched nanoclusters, $n = 119$ PSD-95 nanoclusters, 90 RIM1/2 nanoclusters. **g**, Enrichment indices for RIM1/2, Munc13, and Bsn relative to PSD-95 nanoclusters (filled) and for the opposite direction (open), $n > 260$ nanoclusters, $^*P < 0.05$; $^{**}P < 0.01$, ANOVA on ranks with Dunn's method. **h**, GluA2 enrichment with respect to RIM1/2 nanoclusters, $n = 36$ synapses. Scale bar, 100 nm. All experiments were repeated ≥3 times. Also see Extended Data Fig. 6 and Supplementary Table 2.

Thus, nanodistribution of RIM predicts the local probability of evoked fusion.

For the synapse as a whole, the impact of presynaptic nanoscale organization and confined vesicle sites (Figs 1 and 2) will depend strongly on whether these RIM nanoclusters align with postsynaptic receptor nanoclusters[4]. To assess this, we compared the distribution of PSD-95 over the face of individual synapses to the corresponding distributions of RIM1/2, as the PSD-95 nanoclusters concentrate higher density of receptors[7]. An example synapse, presented in Fig. 3a (Supplementary Video 1), shows three RIM1/2 nanoclusters and three PSD-95 nanoclusters that appear well-aligned and one pair not aligned. We used two independent approaches to assess the relationship between active zone and postsynaptic density (PSD) protein distributions. First we adapted a paired cross-correlation function (PCF) to measure the spatial relationship between the two distributions (see Methods). The measured active zone–PSD distributions showed a significantly elevated PCF compared to simulated active zone–PSD distributions with either distribution fully randomized (Fig. 3b). We then tested the contribution of nanocluster positions to this elevated PCF (Fig. 3c). Randomizing nanocluster positions and out-of-nanocluster molecules (keeping localizations within nanocluster borders intact) abolished the PCF to chance level, while randomizing just the out-of-nanocluster molecules only modestly reduced the PCF, indicating that the precise positioning of the nanoclusters themselves dominate the overall correlation of protein distributions (Fig. 3c, d).

Second, we reasoned that if synapses were trans-synaptically aligned on the nanoscale level, the protein distribution on one side of the synapse would predict protein density in the opposing neuron. To test this, we measured RIM1/2 localization densities as a function of radial distance from the centres of PSD-95 nanoclusters as translated across the synaptic cleft (Fig. 3e). RIM1/2 localization densities within a 60 nm radius were significantly higher than the synaptic cluster average, decaying e-fold per 43.2 ± 12.1 nm away from the peak. This enrichment was again principally dependent on the relative positioning of nanoclusters within synaptic clusters (Fig. 3e). For each individual nanocluster, we defined an enrichment index as the average molecular density of the opposed protein within a 60 nm radius from the nanocluster centre (Extended Data Fig. 7a). Nanoclusters with enrichment indices significantly greater than that of the fully randomized distribution were considered enriched (Fig. 3f). We found 44.4 ± 3.0% of PSD-95 nanoclusters to be enriched (Extended Data Fig. 7b), and these nanoclusters were opposed to RIM1/2 molecule densities that were 2.0 ± 0.1 times the average RIM1/2 synaptic cluster density (Fig. 3f). A similar PSD-95 protein enrichment profile was found relative to the centres of RIM1/2 nanoclusters (Fig. 3e). Thus, this detailed metric for assessing nanoscale alignment revealed strong co-enrichment of these key proteins along narrow, transcellular columns. In comparison to RIM1/2, the enrichment of Munc13 with respect to PSD-95 nanoclusters was considerably weaker, and Bsn intermediate (Fig. 3d, g, Extended Data Fig. 7c–e, Supplementary Table 2). Together, both the PCFs and protein enrichment analyses revealed significant trans-synaptic alignment between RIM1/2 and PSD-95 distributions, largely stemming from the correlated positions of their respective nanoclusters. We likewise found quantitatively similar number, characteristics, and alignment of pre- and postsynaptic nanoclusters in acute hippocampal slices from adult rats (Extended Data Fig. 7f–h).

To determine whether evoked release aligns with postsynaptic receptors, we compared distributions of GluA2-containing AMPARs with RIM1/2 (Fig. 3h). Similar to PSD-95, GluA2 was significantly enriched relative to RIM1/2 nanoclusters, decaying e-fold per 66.9 ± 15.4 nm. This was further confirmed with a different GluA2/3 antibody (Supplementary Table 2). Importantly, given that the probability of AMPAR activation declines with distance from glutamate release sites has previously been deduced[3,16], we can predict synaptic potency by using the observed RIM1/2 and receptor distributions. To estimate the physiological impact of this trans-synaptic alignment, we calculated receptor activation in a measured synapse versus randomized distributions. Consistent with effect sizes posited by previous models[4,5,17], the measured distribution with trans-synaptic alignment gained 21.8 ± 0.5% in synaptic strength compared to a uniform distribution of active zone and PSD proteins (Extended Data Fig. 8), suggesting this synaptic architecture facilitates higher single-vesicle response potency. For comparison, long-term depression induces a very similar magnitude decrease in synaptic strength[18].

Notably, we found that trans-synaptic molecular alignment may extend deeper into the postsynaptic cell, as postsynaptic scaffold molecules farther from the plasma membrane also colocalized with PSD-95 nanoclusters (Extended Data Fig. 9a, c), and RIM1/2 was correspondingly enriched with respect to Shank nanoclusters (Extended Data Fig. 9b). 3D-STORM imaging of RIM1/2, PSD-95, and GKAP1 at the same synapses further confirmed their mutual co-enrichment (Extended Data Fig. 9d–f). Altogether, these results revealed an

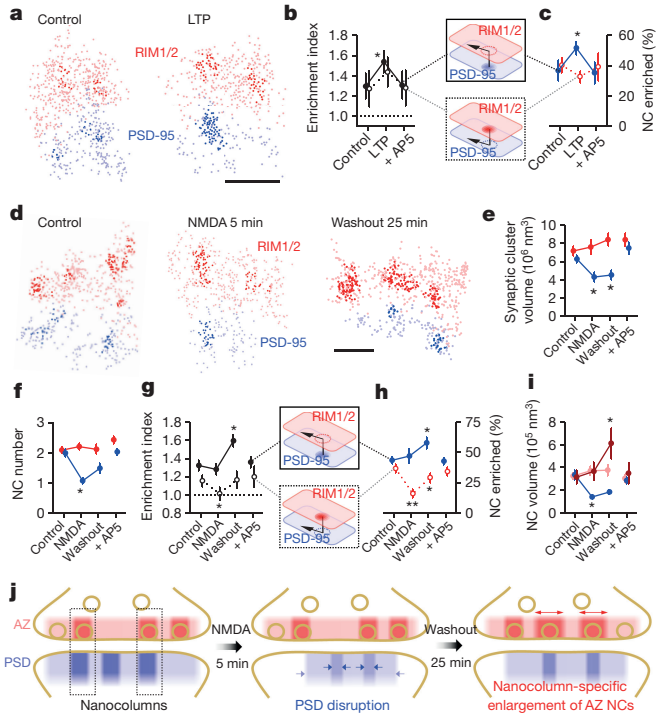


Figure 4 | Retrograde plasticity of synaptic nanoscale alignment. **a**, Distributions of synaptic RIM1/2 and PSD-95 for control and post-LTP induction conditions with nanoclusters highlighted. **b**, **c**, Across-condition comparison of enrichment index and percentage of nanoclusters enriched ($n = 45$, 87 and 42 synapses for control, LTP and AP5, respectively). **d**, Distributions of RIM1/2 and PSD-95 for conditions following NMDA stimulation. Scale bar, 100 nm. **e**–**i**, Across-conditions comparison of RIM1/2 and PSD-95. Dark red in **i** represents RIM1/2 nanoclusters enriched with PSD-95 and light red the unenriched nanoclusters. $n = 61$, 96, 77 and 74 synapses for control, NMDA, washout and AP5, respectively. **j**, Schematic summarizing the reorganization of nanoclusters during NMDA-induced plasticity and recovery. $^*P < 0.05$; $^{**}P < 0.01$, ANOVA on ranks with pairwise comparison (Dunn's method), and χ^2 test for the proportion. All experiments were repeated $\geq$3 times.

axially oriented molecular ensemble spanning the cleft within the bounds of the synapse, evoking the concept of a trans-synaptic nanocolumn enriched with key proteins that regulate synaptic transmission (Extended Data Fig. 9g). The graded protein densities involved suggest this may not be a clearly delineated structural element. Nevertheless, sensitivity of PSD-95 nanocluster size to latrunculin[4] further suggests that the spine cytoskeleton is engaged at the base of the column. Because actin executes many aspects of synaptic plasticity, this provides a potential means by which synaptic strength may be dynamically tuned.

Consequently, we speculated that nanoscale alignment might be altered during synaptic plasticity. To test this, we induced long-term potentiation via glycine stimulation and withdrawal of the NMDAR antagonist D,L-2-amino-5-phosphonovaleric acid (AP5)[19], which resulted in an increase in PSD-95 localization density within nanoclusters, in the percentage of PSD-95 nanoclusters enriched with RIM1/2, and in the enrichment index of PSD-95 nanoclusters (Fig. 4a–c, Extended Data Fig. 10m). These changes were prevented by co-application of AP5 (Fig. 4a–c, Extended Data Fig. 10m). Notably, no changes in RIM1/2 were observed, consistent with LTP as a primarily postsynaptic phenomenon.

We next tested an acute 5-min activation of NMDARs, known to induce a sustained depression of synaptic strength[20,21]. Following this stimulus, postsynaptic nanostructure was markedly disrupted in the generally opposite manner, with the synaptic cluster volume of PSD-95 and the number, volume, and protein density of PSD-95 nanoclusters all reduced (Fig. 4d–f, Supplementary Table 3). These effects were long-lasting, and during the subsequent 25 min, most parameters underwent only partial recovery. In contrast, presynaptic nanostructure underwent a remarkably different pattern of reorganization that was detectable only in relation to PSD-95 nanoclusters. Unlike PSD-95, RIM1/2 distributions were not affected immediately following the stimulus (Fig. 4d–f). However, following the 25-min recovery, the enrichment index of RIM1/2 with respect to PSD-95 nanoclusters increased with a corresponding increase in the percentage of enriched PSD-95 nanoclusters (Fig. 4g, h). Remarkably, while RIM1/2 nanoclusters altogether remained constant in number and enriched percentage, there was in fact an increase in the size of those RIM1/2 nanoclusters that were enriched with PSD-95, whereas the other non-enriched RIM1/2 nanoclusters remained constant (Fig. 4i). Similar results were found when we studied NMDA-induced changes on RIM1/2 and GluA2/3 alignment (Extended Data Fig. 10a–h). Note that on a traditional microscopic level, these changes to presynaptic organization were essentially undetectable: RIM1/2 staining revealed no change in synaptic cluster size or intensity at any point. Because the delayed presynaptic modification was specific to aligned nanoclusters, it may be that nanocolumns point to an alignment-specific, retrograde presynaptic compensation following postsynaptic depression (Fig. 4j), potentially relating to previous reports of presynaptic homeostatic plasticity[22].

Overall, the gradients of protein density we observed suggest a nanocolumn model, in which active zone regions with the highest likelihood of release are aligned to the densest receptor areas, optimizing the potency of neurotransmission (Supplementary Video 2). This provides a simple organizational principle that may hold for many small, central nervous system synapses, and will have the largest influence at synapses that typically release only one vesicle following an action potential. The compartmentalized active zone architecture is reminiscent of protein organization in *Drosophila* neuromuscular junction[23] and vertebrate ribbon synapses, where vesicles and priming proteins are arrayed around tight clusters of Ca^{2+} channels. However, observations in small central nervous system synapses of both clustered[24,25] and random distributions of Ca^{2+} channels[26], and emerging evidence for channel mobility as an equalizer of P_r for vesicles independent of channel positioning[27], suggest that their precise distribution may not be the sole determinant of the active zone release likelihood landscape.

The alignment of pre and postsynaptic nanoscale subdomains[4–6] suggests that even small synapses may be composed of dynamic functional modules[28,29]. We hypothesize that the nanocolumn represents an especially sensitive point whereby disease-associated pathways, frequently known to alter synaptic plasticity[1,2], may disrupt synapse function. It will be important to identify which, if any, of the numerous cleft-spanning adhesion systems[30] or retrograde signalling mechanisms mediate release-receptor alignment and permit dynamic trans-synaptic realignment.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 24 July 2015; accepted 27 June 2016.
Published online 27 July 2016.

1. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506,** 179–184 (2014).
2. Volk, L., Chiu, S.-L., Sharma, K. & Huganir, R. L. Glutamate synapses in human cognitive disorders. *Annu. Rev. Neurosci.* **38,** 127–149 (2015).
3. Franks, K. M., Stevens, C. F. & Sejnowski, T. J. Independent sources of quantal variability at single glutamatergic synapses. *J. Neurosci.* **23,** 3186–3195 (2003).
4. MacGillavry, H. D., Song, Y., Raghavachari, S. & Blanpied, T. A. Nanoscale scaffolding domains within the postsynaptic density concentrate synaptic AMPA receptors. *Neuron* **78,** 615–622 (2013).
5. Nair, D. *et al.* Super-resolution imaging reveals that AMPA receptors inside synapses are dynamically organized in nanodomains regulated by PSD95. *J. Neurosci.* **33,** 13204–13224 (2013).

6. Fukata, Y. *et al.* Local palmitoylation cycles define activity-regulated postsynaptic subdomains. *J. Cell Biol.* **202,** 145–161 (2013).
7. Südhof, T. C. The presynaptic active zone. *Neuron* **75,** 11–25 (2012).
8. Huang, B., Wang, W., Bates, M. & Zhuang, X. Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy. *Science* **319,** 810–813 (2008).
9. Dani, A., Huang, B., Bergan, J., Dulac, C. & Zhuang, X. Superresolution imaging of chemical synapses in the brain. *Neuron* **68,** 843–856 (2010).
10. Park, H., Li, Y. & Tsien, R. W. Influence of synaptic vesicle position on release probability and exocytotic fusion mode. *Science* **335,** 1362–1366 (2012).
11. Watanabe, S. *et al.* Ultrafast endocytosis at mouse hippocampal synapses. *Nature* **504,** 242–247 (2013).
12. Balaji, J. & Ryan, T. A. Single-vesicle imaging reveals that synaptic vesicle exocytosis and endocytosis are coupled by a single stochastic mode. *Proc. Natl Acad. Sci. USA* **104,** 20576–20581 (2007).
13. Leitz, J. & Kavalali, E. T. Fast retrieval and autonomous regulation of single spontaneously recycling synaptic vesicles. *eLife* **3,** e03658 (2014).
14. Betzig, E. Single molecules, cells, and super-resolution optics (Nobel Lecture). *Angew. Chem. Int. Ed.* **54,** 8034–8053 (2015).
15. Levet, F. *et al.* SR-Tesseler: a method to segment and quantify localization-based super-resolution microscopy data. *Nat. Methods* **12,** 1065–1071 (2015).
16. Raghavachari, S. & Lisman, J. E. Properties of quantal transmission at CA1 synapses. *J. Neurophysiol.* **92,** 2456–2467 (2004).
17. Tarusawa, E. *et al.* Input-specific intrasynaptic arrangements of ionotropic glutamate receptors and their impact on postsynaptic responses. *J. Neurosci.* **29,** 12896–12908 (2009).
18. Dudek, S. M. & Bear, M. F. Homosynaptic long-term depression in area CA1 of hippocampus and effects of *N*-methyl-d-aspartate receptor blockade. *Proc. Natl Acad. Sci. USA* **89,** 4363–4367 (1992).
19. Araki, Y., Zeng, M., Zhang, M. & Huganir, R. L. Rapid dispersion of SynGAP from synaptic spines triggers AMPA receptor insertion and spine enlargement during LTP. *Neuron* **85,** 173–189 (2015).
20. Lee, H.-K., Kameyama, K., Huganir, R. L. & Bear, M. F. NMDA induces long-term synaptic depression and dephosphorylation of the GluR1 subunit of AMPA receptors in hippocampus. *Neuron* **21,** 1151–1162 (1998).
21. Sanderson, J. L. *et al.* AKAP150-anchored calcineurin regulates synaptic plasticity by limiting synaptic incorporation of Ca^{2+}-permeable AMPA receptors. *J. Neurosci.* **32,** 15036–15052 (2012).
22. Davis, G. W. & Müller, M. Homeostatic control of presynaptic neurotransmitter release. *Annu. Rev. Physiol.* **77,** 251–270 (2015).
23. Liu, K. S. *et al.* RIM-binding protein, a central part of the active zone, is essential for neurotransmitter release. *Science* **334,** 1565–1569 (2011).
24. Holderith, N. *et al.* Release probability of hippocampal glutamatergic terminals scales with the size of the active zone. *Nat. Neurosci.* **15,** 988–997 (2012).
25. Nakamura, Y. *et al.* Nanoscale distribution of presynaptic Ca^{2+} channels and its impact on vesicular release during development. *Neuron* **85,** 145–158 (2015).
26. Scimemi, A. & Diamond, J. S. The number and organization of Ca^{2+} channels in the active zone shapes neurotransmitter release from Schaffer collateral synapses. *J. Neurosci.* **32,** 18157–18176 (2012).
27. Schneider, R. *et al.* Mobility of calcium channels in the presynaptic membrane. *Neuron* **86,** 672–679 (2015).
28. Tarr, T. B., Dittrich, M. & Meriney, S. D. Are unreliable release mechanisms conserved from NMJ to CNS? *Trends Neurosci.* **36,** 14–22 (2013).
29. Lisman, J. & Raghavachari, S. A unified model of the presynaptic and postsynaptic changes during LTP at CA1 synapses. *Sci. STKE* **2006**, re11 (2006).
30. Missler, M., Südhof, T. C. & Biederer, T. Synaptic cell adhesion. *Cold Spring Harb. Perspect. Biol.* **4,** a005694 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank S. Thompson, T. Abrams, S. Jurado and G. Wittenberg for advice and comments, P. Kaeser for advice on RIM expression and RIM antibodies, Y. Araki and R. Huganir for advice on chemLTP, and S. S. Divakaruni for advice and initial tests of chemLTP. We thank P. Kaeser for the gift of RIM1-mVenus, T. Ryan for vGlut1-pHluorin-mCherry, G. Augustine for Syn1a–CFP, and M. Contreras for technical assistance. This work was supported by F30-MH105111 to H.C., F30-MH102891 to T.P.L., F31-MH105105 to S.R.M., T32-GM008181 to H.C. and S.R.M., R01-MH080046 and NS090644 to T.A.B, and a gift from the Kahlert Foundation to T.A.B.

Author Contributions A.T. and H.C. performed STORM experiments, A.T. designed 3D-STORM analysis, H.C. performed and analysed pHuse and RIM PALM experiments, T.P.L. and A.T. performed simulations, S.R.M. performed GCaMP imaging and nanobody STORM experiments, H.D.M. performed PSD PALM experiments, and A.T., H.C. and T.A.B. designed the experiments and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.A.B. (tblanpied@som.umaryland.edu) or A.T. (tangaihui@gmail.com).

Reviewer Information *Nature* thanks S. Sigrist, X. Zhuang and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

All experimental protocols were approved by the University of Maryland, Baltimore School of Medicine Institutional Animal Care and Use Committee. Dissociated hippocampal neurons from E18 SD rats of both sexes were prepared as described previously[31]. To increase the experiment efficiency, for three-colour STORM experiments we used the 'sandwich' cultures with a supporting astroglial monolayer as described previously[32] in which most neuronal structures were in the same focal plane. All experiments were performed on neurons 14–21 DIV and repeated on 3 or more separate cultures unless otherwise specified.

Immunostaining. Cells were fixed with 4% paraformaldehyde (PFA) and 4% sucrose in PBS (pH 7.4) for 10 min at room temperature (RT), followed by washing with 50 mM glycine in PBS. Cells were then permeabilized and blocked using 3% BSA or 5–10% donkey or goat serum in PBS with 0.1% Triton X-100, followed by incubation with primary antibody (3 h RT or 4 °C overnight) and secondary antibodies (1 h RT).

For comparisons of Munc13 or RIM1/2 with Bsn made using 3D-STORM, mouse anti-Bsn (1:500, Enzo) was used with either rabbit anti-RIM1/2 (1:500; Synaptic Systems No. 140203) or rabbit anti-Munc13 (1:500; Synaptic Systems No. 126103). Cy3 or Alexa-647 conjugated goat or donkey anti-rabbit or anti-mouse secondary antibodies (1:200 in PBS; JacksonImmuno) were used[33]. For comparisons of Munc13 and RIM1/2, staining was performed sequentially separated by additional blocking steps of incubation with rabbit serum at RT for 30 min followed by incubation with excess unconjugated anti-rabbit Fab antibody for 1 h at RT. For this set of experiments, all permutations of the order in which the primary antibody was applied and the fluorophore used to label each protein were included. For trans-synaptic measurements, rabbit anti-Munc13, anti-RIM1/2, anti-RIM1 (1:500; Synaptic Systems No.140003) or anti-Bsn (1:500, Cell Signaling), were used with mouse anti-PSD-95 (1:200; Neuromab), mouse anti-GluA2 (1:100, Millipore), or rabbit anti-GluR2/3 (1:100, Millipore). Unless specified otherwise, presynaptic proteins were labelled with donkey anti-rabbit IgG conjugated with Alexa-647 and postsynaptic PSD-95 were labelled with donkey anti-mouse IgG conjugated with Cy3. For comparison of directly labelled primary antibody with primary–secondary antibody labelling, we directly conjugated Alexa-647 dye to anti-PSD-95 antibody and purified antibody using illustra NAP Columns (GE Healthcare). For comparison of nanobody labelling of expressed GFP-tagged knockdown-rescue PSD-95 with primary–secondary antibody labelling, we used GFP-booster (1:200, Chromotek). More information on antibodies used can be found in the Supplementary Information.

Tissue slice staining was performed essentially as previously described[9,34]. Briefly, 1-mm thick blocks of hippocampal tissue from 5–7-week-old male SD rats were fixed with ice-cold 4% PFA for 15 min and then dehydrated with 30% sucrose in PBS. Cryostat sections with 40 μm thickness were made, permeabilized and blocked with 10% donkey serum and 0.3% Triton X-100 in PBS/glycine for 1 h. PSD-95 and RIM1/2 were labelled with the same antibody concentration as was used in cell culture.

3D-STORM imaging. Imaging was performed on an Olympus IX81 ZDC2 inverted microscope with a 100×/1.49 TIRF oil-immersion objective. Excitation light was reflected to the sample via a 405/488/561/638 quad-band polychroic (Chroma). The typical incident power was ~30 mW for 647 nm and ~60 mW for 561 nm (measured through the objective). To reduce background fluorescence while maximizing the depth of view, we adjusted the incident angle of the excitation beam to near but less than the critical angle, to achieve oblique illumination of the sample. Emission was passed through a Photometrics DV2 which split the emission at 565 nm and directed the red and far-red bands through matched filters (595/50 and 655 long-pass) onto an iXon+ 897 EM-CCD camera (Andor). A cylindrical lens (focal length = 30 cm) was inserted in each path of the splitting cassette of the DV2 to create the astigmatism for 3D imaging. All hardware was controlled via iQ software (Andor). *Z* stability was maintained by the Olympus ZDC2 feedback positioning system. Lateral drift was corrected with a cross-correlation drift-correction approach[35,36].

Labelled cells and tissue slices were imaged in a STORM imaging buffer freshly made before experiments containing 50 mM Tris, 10 mM NaCl, 10% glucose, 0.5 mg/ml glucose oxidase (Sigma), 40 μg/ml catalase (Sigma), and 0.1 M cysteamine (Sigma). For tissue slices, the focal plane was set to within 1.5 μm from the glass coverslip to obtain the best signal-to-noise ratio. Imaging was performed as previously described[4,33]. TetraSpeck beads (100 nm; Invitrogen) deposited on a coverslip were localized to correct alignment between the two channels as described previously[4]. The average deviation of the bead localizations after correction was between 10 and 15 nm. To calibrate the 3D positions of localizations, a *z*-stack with 30-nm steps was collected on the same coverslip with beads. The average deviation of localized *z*-positions of immobilized fluorophores was 40–50 nm.

Three-colour 3D-STORM. Three-colour STORM were performed with two sequential sets of two-colour 3D-STORM on RIM1/2-PSD-95 as a pair and then GKAP1-PSD-95 as a pair. Cells were immunolabelled with mouse anti-PSD-95, rabbit anti-RIM1/2, and mouse anti-GKAP1 (1:200, Neuromab). PSD-95 and RIM1/2 were then immunolabelled with secondary antibodies conjugated to Alexa647 and Cy3, respectively. After >20 min of continuous excitation by high-powered lasers during the first round of imaging, the majority of Cy3 molecules (RIM1/2) became bleached. After acquisition of the first set of data, GKAP1 was then labelled with secondary antibody conjugated to Cy3 while the coverslip remained on the microscope. The two sets of data were aligned post hoc based on Alexa647 (PSD-95) localizations. Because RIM1/2 and GKAP1 are not overlapping proteins, in the second imaging set, those Cy3 localizations within the RIM cluster borders potentially arising from the small, unbleached fraction of RIM-Cy3 were rejected from GKAP1 localizations.

PALM-STORM Imaging. PALM imaging of PSD-95 concurrent with STORM imaging of GKAP or Shank (1:200, Neuromab) was performed as previously described[4].

Single-molecule localization and analysis. All data analysis was performed offline using custom routines in MATLAB (Mathworks). Molecule locations were determined by fitting an elliptical 2D Gaussian function to an 11 × 11 pixel array (pixel size 100 nm) surrounding the peak[4]. The lateral (x, y) and axial (z) coordinates of the fluorophore were determined from the centroid position and ellipticity of the fitted peak, respectively[8]. Only molecules localized with an x–y precision <10 nm (ref. 37), fitting $R^2 > 0.6$, and comprising >200 photons were used for further analysis.

To remove the localizations from those fittings of multiple overlapping peaks, we developed a rejection criteria based on the shape of peaks. For peaks arising from single fluorophores, the fitted width in x and y (W_x and W_y, respectively) should correlate in a manner mainly determined by the cylindrical lens. All localizations away from this correlation would come from multiple overlapping or poorly fitted peaks and were therefore rejected (Extended Data Fig. 1a–f).

Single-molecule tracking was employed to remove the overcounted localizations from peaks lasting for more than one frame. Tracking was accomplished with available algorithms (http://physics.georgetown.edu/matlab/). Particles appearing in consecutive frames separated by no more than 200 nm were collapsed into one track and considered one molecule by taking only the location in the first frame for further analysis.

Analysis of synaptic clusters. A potential synapse could be identified by a juxtaposed pair of synaptic proteins in a 2D scatter plot of all accepted localizations from both channels. By rotating a 3D scatter plot of localizations of a selected potential synapse, we evaluated the data quality and selected only those with clear pre- and postsynaptic components (for example, no nearby third cluster which may indicate two synapses in close proximity) for further analysis. To define the border of a synaptic cluster, the nearest neighbour distances (NND) between localizations were calculated and the mean + 2 s.d. of NND was used as a cut-off to divide the localizations into sub-clusters. All localizations outside of the primary sub-clusters were considered to be background and discarded.

Owing to the irregularly curved shapes of some synapses, using the convex hull to define synaptic cluster shape would overestimate the synaptic cluster volume. We thus defined the synaptic cluster using the alpha shape of the set of 3D localizations with $\alpha = 150$ nm. This value was determined based on series of tests on >100 synapses to obtain the best synaptic cluster shape while avoiding dramatic changes in volume when individual localizations near the border were added or removed. This alpha shape algorithm gave a synaptic cluster volume of 81 ± 3% of the convex hull volume ($n = 156$ synapses). Subsequently, this alpha shape was used as the cluster border when localizations were randomized.

A synaptic cluster was only considered for analysis if the volume was between 2×10^{-3} μm^3 and 30×10^{-3} μm^3 (ref. 38), and contained an average density of $>8 \times 10^3$ localizations/μm^3. Local density was defined as the number of molecules within a radius of 2.5 times the standard median nearest neighbour distance (MdNND) for the synaptic cluster density. The standard MdNND was calculated from a standard correlation curve $\mathrm{MdNND} = \sqrt[3]{\frac{0.174}{d}}$ (unit per 100 nm voxel for d) where d is the averaged localization density. This equation is derived from fitting MdNND with d in a series of simulations of uniformly distributed synaptic clusters with different densities. The reason we used this standard MdNND instead of the median NND from the original synaptic cluster was to reduce the deviation caused by local assembles.

Nanocluster analysis. Localizations with local densities ≥14 were selected and divided into agglomerative sub-clusters with a node height cut-off of 40 nm using MATLAB functions linkage() and cluster(). For each sub-cluster, we then calculated the NND and discarded those localizations with NND > MdMND if any. Only those sub-clusters containing ≥4 localizations were counted as nanoclusters.

These criteria were chosen based on a conservative strategy such that no nanoclusters were identified in simulations of randomly distributed synaptic clusters

with different densities. Consequently, they may have prevented detection of small or weakly enriched nanoclusters. In principal, we cannot completely exclude the possibility of overcounting, so a certain fraction of detected nanoclusters are potentially artificial. However, we used the same standard on all data sets. Since all the trans-synaptic analyses were well controlled by randomizing simulations, this contamination is not able to produce false positives for trans-synaptic alignment analyses. On the contrary, it may attenuate the significance of the differences in trans-synaptic analyses based on nanoclusters, including cross-correlation, protein enrichment and the fraction of enriched nanoclusters.

Since the number of localizations in one nanocluster was typically small, using convex hull or alpha shape would greatly under-estimate the nanocluster volume due to the border effect. Therefore, we tessellated the synaptic cluster with polyhedrons using MATLAB function voronoin(), with each Voroni cell containing one localization. The nanocluster volume was calculated as a summation of volumes of all polyhedrons containing the nanocluster localizations. To avoid unexpected unbounded Voronoi cells and over-estimating the volume of cells near the cluster surface, we introduced ~10% background noise by adding randomly distributed localizations around the cluster[15]. Polyhedron volume for each localization was averaged across ten independent simulations.

ACF analysis. To quantify the self-clustering of synaptic proteins, we adapted an autocorrelation function[4,39] for our 3D data. The autocorrelation function $g_a(r)$ is a measure of density correlations, which reports increased probability of finding a second localized signal a distance r away from a given localized signal. It was tabulated in Matlab using Fast Fourier Transforms (FFTs), as in equation (1).

$$g_a(\vec{r}) = \frac{\mathrm{FFT}^{-1}(|\mathrm{FFT}(\boldsymbol{I})|^2)}{\rho^2 \mathrm{FFT}^{-1}(|\mathrm{FFT}(W)|^2)} \quad (1)$$

FFT^{-1} is an inverse Fast Fourier Transform, $\boldsymbol{I}$ is the reconstructed 3D density matrix of localized fluorophores (pixel size of 5 nm), ρ is the general localization density inside the synaptic cluster, and W is a shape function that has the value of 1 inside the synaptic cluster as defined above with an alpha shape and the value of 0 elsewhere. The matrix $\boldsymbol{I}$ was padded with zeros in all three directions out to a distance larger than the range of the desired correlation function (we used 200 nm) to avoid artefacts due to the periodic nature of FFT functions. W was also padded by an equal number of zeros. $\mathrm{FFT}^{-1}(|\mathrm{FFT}(W)|^2)$ is a normalization factor accounting for the general shape of the synaptic cluster itself so that the output of the $g_a(\vec{r})$ represented only the internal structure of the measured synaptic cluster. $g_a(\vec{r})$ was symmetric to rotations around the centre of matrix $\boldsymbol{C}$ (x_c, y_c, z_c), and it could be averaged over angles to obtain $g_a(r)$ by converting to polar coordinates. $g_a(r)$ was then binned by radius (r). Correlation functions were plotted for $r > 0$, as $g_a(r = 0)$ was a trivial contribution.

For a uniform distribution, for example, when all localizations were uniformly randomized within the alpha shape, $g_a(r) = 1$ (Fig. 1d). Any heterogeneity will result in a $g_a(r) > 1$. The extent of $g_a(r)$ over 1, that is, r_0 for $g_a(r_0) = 1$, is related to the pattern size of the internal heterogeneity (Extended Data Fig. 2b, c)[39].

Isolated, non-synaptic small groups of localizations were taken from our experimental data. These localization groups likely represent an overestimate of a single-dye-molecule localization spread. Nevertheless, we find that they are still significantly smaller than the large majority of the nanoclusters we detected.

Imaging vesicle exocytosis. For imaging vesicle fusion, vGluT-pHluorin-mCherry (a gift from T. Ryan)[40,41], was cotransfected with Syn1a–CFP (a gift from G. Augustine) using Lipofectamine 2000 (Invitrogen) for 4–6 days before imaging cells at 14–20 DIV. Optical measurements were performed using a laminar-flow perfusion and stimulation chamber. Images were acquired at 10 Hz with an Andor iXon 887 EM-CCD camera on an Olympus IX81 ZDC2 inverted microscope with a 100×/1.49 TIRF oil-immersion objective. Temperature was controlled using an objective heater set at either room temperature (~25 °C) or 32 °C. Action potentials were evoked by passing 1 ms current pulses yielding fields of ≈10 V/cm via platinum-iridium electrodes. Terminals were selected for imaging by assessing their responsiveness, as indicated by a fluorescence increase, to a 10 AP train at 20 Hz. A wide-field Syn1a image was then taken at the imaging plane. Single AP-evoked release was measured over 60 trials of (1) 1 s acquisition of baseline fluorescence, (2) stimulus, (3) 2.5 s acquisition of post-stimulus fluorescence, (4) 7 s recovery during which the laser is off. Spontaneous release was measured over 5 min of continuous acquisition. Cells were imaged in a saline solution containing 120 mM NaCl, 3 mM KCl, 2 mM $CaCl_2$, 2 mM $MgCl_2$, 10 mM glucose, and 10 mM HEPES, pH adjusted to 7.4 with NaOH, 10 μM 6,7-dinitroquinoxaline-2,3-dione (DNQX; Sigma), 50 μM D,L-2-amino-5-phosphonovaleric acid (AP5; Sigma), and 500 nM Jasplikinolide (Jasp; Millipore) at room temperature (~25 °C). When higher [$CaCl_2$] was used, [$MgCl_2$] was reduced to keep the divalent ion concentration constant. For measurements of spontaneous events, 500 nM tetrodotoxin (TTX; Enzo) was added after identifying terminals using AP-evoked fluorescence increase.

For calculating normalized changes in fluorescence ($\Delta F/F$), images were analysed in ImageJ by custom-written plugins[12]. Average fluorescence intensities were measured over a circular region of interest (ROI) of radius 800 nm for each bouton. Change in fluorescence (ΔF) was calculated as the difference in intensity of the frame after the stimulus was delivered and the average ROI intensity of 5 baseline frames not including the first frame or the frame immediately before the stimulus (F_{baseline}). $\Delta F/F$ was calculated by normalizing each ΔF to F_{baseline}.

pHuse localization and analysis. Data analysis was performed offline using custom routines in MATLAB (Mathworks). Boundaries for individual boutons were determined using wide-field images of Syn1a–CFP centred at the focal plane of the pHuse experiments thresholded at 50% of the peak intensity (33% and 67% thresholds were also compared and showed no significant difference on the effect of mode of release, shown in Extended Data Fig. 5i). Binary images were created from the thresholded image, and Syn1a–CFP puncta area calculated as a measure of bouton area, which correlated with pHuse area, as expected[42]. Images for each fusion event were processed using frame-by-frame subtraction followed by background subtraction to isolate fluorescence increases (Fig. 2d)[43]. Similar detection thresholds were set for spontaneous (75 ± 15) and evoked (78 ± 14, $t = 0.88$, $P = 0.40$) release, at ~3–4 times above background noise, on an individual imaging field basis. Spatial localization of the fusion events was determined by fitting an elliptical 2-dimensional Gaussian function to a 9×9 pixel array surrounding the peak. Only molecules localized with a precision <25 nm[37,44], elliptical form <1.3, and comprising >100 photons were used for further analysis. An additional criterion to exclude evoked pHuse localizations with photon counts $>$ mean + 2 s.d. of spontaneous photon count distribution was used in Extended Data Fig. 5d and showed no significant difference compared to the distribution lacking this criterion. Localizations from multiple fusion events over time at individual boutons were mapped. A 2D convex hull algorithm was used to calculate the minimal convex polygon that incorporated all fusion site localization points. The area of the resulting polygon was used as the fusion site (pHuse) area.

Photon count distributions analysis. Data analysis was performed offline using custom routines in MATLAB (Mathworks). The distribution for spontaneous fusion events was fit with a normal distribution using normfit(), which uses maximal likelihood estimation for optimization. The distribution of evoked fusion events was fit with a custom univariate distribution for a mixture of two normal distributions with a probability density function (pdf) defined in equation (2). This fitting also used maximal likelihood estimation for optimization of five parameters, including the mixture probability (p), and the population means (μ_1, μ_1) and variance (σ_1, σ_2) for each component, over 300 iterations using normpdf() to compute the pdf for each of the two component normal distributions.

$$\mathrm{pdf} = p \times \mathrm{normpdf}(x, \mu_1, \sigma_1) + (1 - p) \times \mathrm{normpdf}(x, \mu_2, \sigma_2) \quad (2)$$

Here p was constrained between 0 and 1, and σ had a lower bound of 0. This mixture probability defined the lower estimate (72%) for the percentage of single stimulus evoked fusion arising from single vesicles. We calculated the higher estimate (82%) by calculating the percentage of evoked fusion events with photon counts within two standard deviations of the mean spontaneous fusion event photon count. To assess the influence of multivesicular events on evoked pHuse area, we used this as a cut-off to exclude localizations above this photon count. We found no significant difference between evoked area with and without excluding these events (Extended Data Fig. 5d).

Ca^{2+} imaging and analysis. For Ca^{2+} imaging, the genetically encoded indicator GCaMP6f (ref. 45) was transfected at 14 DIV and imaged 3 days after transfection. GCaMP6f was used to detect postsynaptic miniature spontaneous Ca^{2+} transients (mSCaTs) that arose in dendritic spines following NMDA receptor activation by spontaneous release[46]. Coverslips were placed in custom-made chambers in saline solution containing 1 μM TTX, 10 μM DNQX, 25 μM picrotoxin (Sigma), and 5 μM nifedipine (Sigma). Imaging was performed on a spinning disk confocal system (Andor Technology), consisting of a CSU-22 confocal (Yokagawa) with a Zyla 4.2 CCD camera detector (Andor) mounted on the side port of an Olympus IX-81 inverted microscope, using a 60×/1.42 oil-immersion objective, yielding a final effective pixel size of 108 nm. Continuous acquisition at 20 Hz was collected for 3 min, controlled by iQ software (Andor).

Data analysis was performed offline using custom routines in Metamorph (Molecular Devices), Clampex (Molecular Devices), and Matlab (Mathworks). First, using Metamorph, a baseline image was created by averaging the first three and last three image frames and a maximum intensity projection was made by averaging all image frames. Image subtraction of the baseline from the maximum intensity projection revealed spines that showed an increase in GCaMP intensity. Regions of interest (ROIs) were drawn around these 'active' spines as well

as a background region and then transferred to the original timelapse. For each ROI the averaged intensity was measured per frame. The average intensity of the background ROI was subtracted from the average intensity of 'active' spine ROIs. From this, an average fluorescence intensity was calculated for every 10 frames, and within every minute interval of imaging the lowest positive value was used as the baseline fluorescence intensity for that minute ($F_{\text{baseline,1 min}}$). A normalized change in fluorescence ($\Delta F/F$) was calculated for each frame as $(F_{\text{frame}} - F_{\text{baseline,1 min}})/F_{\text{baseline,1 min}}$. The $\Delta F/F$ values were then fed into Clampex, and mSCaTs were detected using a template search that identified peaks based on a shape profile determined from mSCaT examples with near-average rise and decay time courses.

Confocal imaging of presynaptic proteins. Neurons 14–20 DIV were cotransfected for 3 days with RIM1-mVenus (a gift from P. Kaeser) and Syn1a–CFP to assess colocalization. Neurons transfected with only RIM1-mVenus were immunostained with chicken anti-GFP (1:200, Chemicon) labelled with secondary anti-chicken-Alexa-488, rabbit anti-RIM1/2 labelled with secondary anti-rabbit-Cy3, and mouse anti-Bsn labelled with secondary anti-mouse-Alexa-647 to assess expression levels. Imaging was performed on a spinning disk confocal system as described above. ImageJ was used to analyse fluorescence intensity of RIM1/2 and Bsn at transfected compared to neighbouring untransfected boutons.

PALM-pHuse. RIM1-mEos3.1 was constructed by subcloning mEos3.1 from mEos3.1-N1 (a gift from S. McKinney) into pCMV5-RIM1-mVenus (P. Kaeser) in place of mVenus at NotI-AscI. PALM was performed on RIM1, and nanoclusters identified using local density measured by nearest neighbour distance as previously described[4], or using an adapted form of SR-Tesseler first rank neighbour density (δ^1), using 2× mean δ^1 of the whole synapse as the threshold for identifying nanoclusters, as described in ref. 15. Nanoclusters identified by both methods were similar in size (Extended Data Fig. 6). To map vesicle fusion to active zone nanostructure, RIM1-mEos3.1 was cotransfected with vGpH at 10–14 DIV and imaged at 14–18 DIV. RIM1 PALM and pHuse of 1-AP-evoked release was performed as described above sequentially on the same boutons. Overlapping RIM1 and pHuse localizations were analysed at boutons containing >10 RIM1 localizations and >3 pHuse localizations offline using custom routines in MATLAB (Mathworks). vGpH fluorescence increase following a 10 AP-train stimulus was used to outline the border of individual boutons. Randomized distributions of RIM1 were simulated for each synapse by randomly placing the same number of RIM1 localizations within the same area of RIM1 as calculated by convex hull of the measured RIM1 distribution. RIM1 local density within these randomized distributions was similarly calculated. Normalized RIM1 δ^1 was calculated with respect to overall synaptic localization density.

3D paired cross-correlation function (PCF) analysis. The 3D PCF was adapted from a similar function previously used to quantify colocalization in 2D data[39]. It was computed using two matrices ($\boldsymbol{I}_1$ and $\boldsymbol{I}_2$) reconstructed from two image channels (equation (3).

$$g_c(\vec{r}) = \mathrm{Re}\left\{\frac{\mathrm{FFT}^{-1}(FFT(\boldsymbol{I}_1) \times \mathrm{conj}[\mathrm{FFT}(\boldsymbol{I}_2)])}{\rho_1\rho_2 \mathrm{FFT}^{-1}(\mathrm{FFT}(W_1) \times \mathrm{conj}[\mathrm{FFT}(W_2)])}\right\} \quad (3)$$

Here, conj[] is a complex conjugate, ρ_1 and ρ_2 are the averaged localization densities in the pair of synaptic clusters, W_1 and W_2 are shape functions of the two synaptic clusters, and Re{} indicates the real part. Different from the ACF, the symmetric origin of $g_c(\vec{r})$ here is no longer the matrix centre $C(x_c, y_c, z_c)$, but a different point $A(x, y, z)$, and the vector $\overrightarrow{CA}$ represents the direction and distance for the translation of PSD-95 synaptic clusters ($\boldsymbol{I}_2$) to get the best overlap with presynaptic clusters ($\boldsymbol{I}_1$). We computed the direct correlation between $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ with equation (4).

$$G = \mathrm{FFT}^{-1}(\mathrm{FFT}(\boldsymbol{I}_1') \times \mathrm{conj}[\mathrm{FFT}(\boldsymbol{I}_2')]) \quad (4)$$

A is the point with the peak G value. Because the originally constructed matrices $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ were not continuous, to reduce the noise of the correlation, we first convoluted the two matrixes with an $11 \times 11 \times 11$ kernel (Extended Data Fig. 1g). To avoid having the correlation be dominated by local domains with high localization density, we cut the peaks of the convoluted matrixes to 1/4 of the mean localization density within synaptic clusters ($\rho_1/4$ and $\rho_2/4$) (Extended Data Fig. 1h) so that G only represented the relationship between the general 3D shapes of the two synaptic clusters ($\boldsymbol{I}_1', \boldsymbol{I}_2'$) without internal heterogeneity (Extended Data Fig. 1m, n). Around A, $g_c(\vec{r})$ is symmetric and could be angularly averaged to get $g_c(r)$.

Since the information of synaptic cluster shape and overall density had been normalized, $g_c(r)$ was fully dependent on the internal organizations of the two synaptic clusters. If localization assemblies inside the two synaptic clusters organized in a similar pattern and opposed each other, $g_c(r) > 1$. If either synaptic cluster had a uniform distribution of localizations (Fig. 3b) or the internal assemblies were not aligned (Fig. 3c), $g_c(r) = 1$. Different from the ACF, overcounting has no effect on the PCF[39].

Protein enrichment analysis. The protein enrichment profile of protein A relative to a protein-B nanocluster, $E_{A\to B}(r)$, was calculated as the angularly averaged localization density of protein A around the aligned centre of a protein-B nanocluster normalized to the average localization density in synaptic cluster A. The aligned nanocluster centre was found as shown in Extended Data Fig. 1. To avoid potential problems caused by boundary conditions, we calculated the enrichment profile as equation (5).

$$E_{A\to B}(r) = \frac{N_{A\to B}(r)}{N_{A_r(m)\to B}(r) \times m} \quad (5)$$

$N_{A\to B}(r)$ is the binned distribution of protein-A localization number to the aligned protein-B nanocluster centre, $N_{A_r(m)\to B}(r)$ is the distribution of localization number for a uniformly randomized synaptic cluster A with m times of original localization density, and m is a factor set to 15 to reduce the effect of fluctuations. A protein-B nanocluster was considered to be significantly enriched with protein A if $E_{A\to B}(r) > \mathrm{mean}[E_{A_r\to B}(r)] + 1.96 \times \mathrm{standarddeviation}[E_{A_r\to B}(r)]$, where $E_{A_r\to B}(r)$ represents the enrichment profile of ten simulated uniformly randomized A synaptic clusters with the original density and the same alignment to the nanocluster centre of protein-B.

Chemical LTP and LTD. Chemical LTP was performed using a combination of AP5 withdrawal and application of glycine as described in ref. 19. Briefly, 3–4-week-old cultures were treated with 200 μM DL-AP5 in culture medium for two days and then transferred to ACSF (150 NaCl, 3 KCl, 2 $CaCl_2$, 1 $MgCl_2$, 10 HEPES-Na, 10 D-glucose, all in mM, pH 7.4) with 100 μM picrotoxin, 1 μM strychnine, 0.5 μM TTX and 200 μM AP5. After preincubation for 1–2 h, chemical LTP was induced with 15 min incubation in the similar solution with 200 μM glycine but without Mg^{2+} and AP5. Neurons were fixed directly following induction. Chemical LTD was performed using application of NMDA as described in ref. 20. Control solutions of regular saline solution or co-application with AP5 were paired with experimental conditions. Cells were fixed either immediately after plasticity induction or washed with saline and incubated for 25 min at 37 °C to allow recovery before fixing. Cells were then immunostained and imaged as described above.

Synaptic modelling. We used an experimentally constrained deterministic approach to study the dependence of synaptic strength on the spatial distribution of release sites and AMPARs. Central to this approach is the relationship between channel opening probability and its distance from a release site, determined previously by stochastic modelling approaches[3,16,47]:

$$P_o(r) = 0.42\ \mathrm{e}^{-r/88} \quad (6)$$

where r is the lateral distance between an AMPAR and a release site (in nm). In brief, the distribution of RIM1/2 proteins and GluA2/3-containing AMPA receptors measured by STORM were used to determine the spatial coordinates of release sites and AMPARs on a model synapse. Since the precise photophysics and blink distribution of dyes are complicated and the exact efficiency of antibody labelling is unknown, we calculated gradient maps of spatial coordinates to determine putative RIM1/2 protein and AMPAR locations from the single-molecule images. First, the 3D spatial coordinates were projected onto 2D planes orthogonal to the manually determined axodendritic axis. Each projected point was assigned a Gaussian function, the amplitude and width of which were determined by the normalized local density and the lateral STORM localization precision (20 nm). Overlapping Gaussian functions within the active zone or PSD convex hull were integrated to create the pre- and postsynaptic gradient maps. The sampling pixel size was 2.5 nm (the calculated synaptic response was independent of pixilation level for sampling size from 1 to 20 nm, data not shown). The pre- and postsynaptic gradient maps were separated by 20 nm, the cleft distance used to determine equation (6)[3].

The model synaptic response for a single synapse was computed as the expected fraction of receptors that would open given a single release, averaged over all possible release locations in the active zone. For any single release event, the expected open fraction of channels at the peak of the response was calculated as follows:

$$O(i) = \sum_j \left[P_o(r_{ij}) \frac{\mathrm{LD}_j}{\sum_j \mathrm{LD}_j}\right] \quad (7)$$

where r_{ij} is the lateral distance between the ith pixel in the presynaptic gradient map and the jth pixel in the postsynaptic gradient map; the expected fraction of open channels $O(i)$ from the ith release site is sum of channel opening probabilities at all pixels in the postsynaptic gradient map, where each jth pixel is weighted by

its normalized local density LD_j (that is, the channel fraction is assumed to be directly proportional to the channel local density). To constrain the location of release events in the active zone, we used the live-cell pHuse-PALM data, which showed that release events preferentially occurred in regions with normalized RIM local density greater than 1.5, and these events occurred over 20–60% of the active zone area (spontaneous pHuse area/PALMed RIM area, and evoked pHuse area/spontaneous pHuse area). To account for these measured features, we modelled the spatial likelihood of release as a piecewise sigmoidal function dependent on the normalized local RIM density:

$$P_r(i\,|\,\text{release}) = \begin{Bmatrix} 0.5\left[\dfrac{(1-s)^{\frac{LD_i}{LD_{inflect}}}}{2-s-\frac{LD_i}{LD_{inflect}}}\right] & \text{if } LD_i \in [0, LD_{inflect}] \\ 0.5+0.5\left[\dfrac{(1-s)^{\frac{LD_i-LD_{inflect}}{LD_{max}-LD_{inflect}}}}{2-s-\frac{LD_i-LD_{inflect}}{LD_{max}-LD_{inflect}}}\right] & \text{if } LD_i \in [LD_{inflect}, LD_{max}] \end{Bmatrix} \quad (8)$$

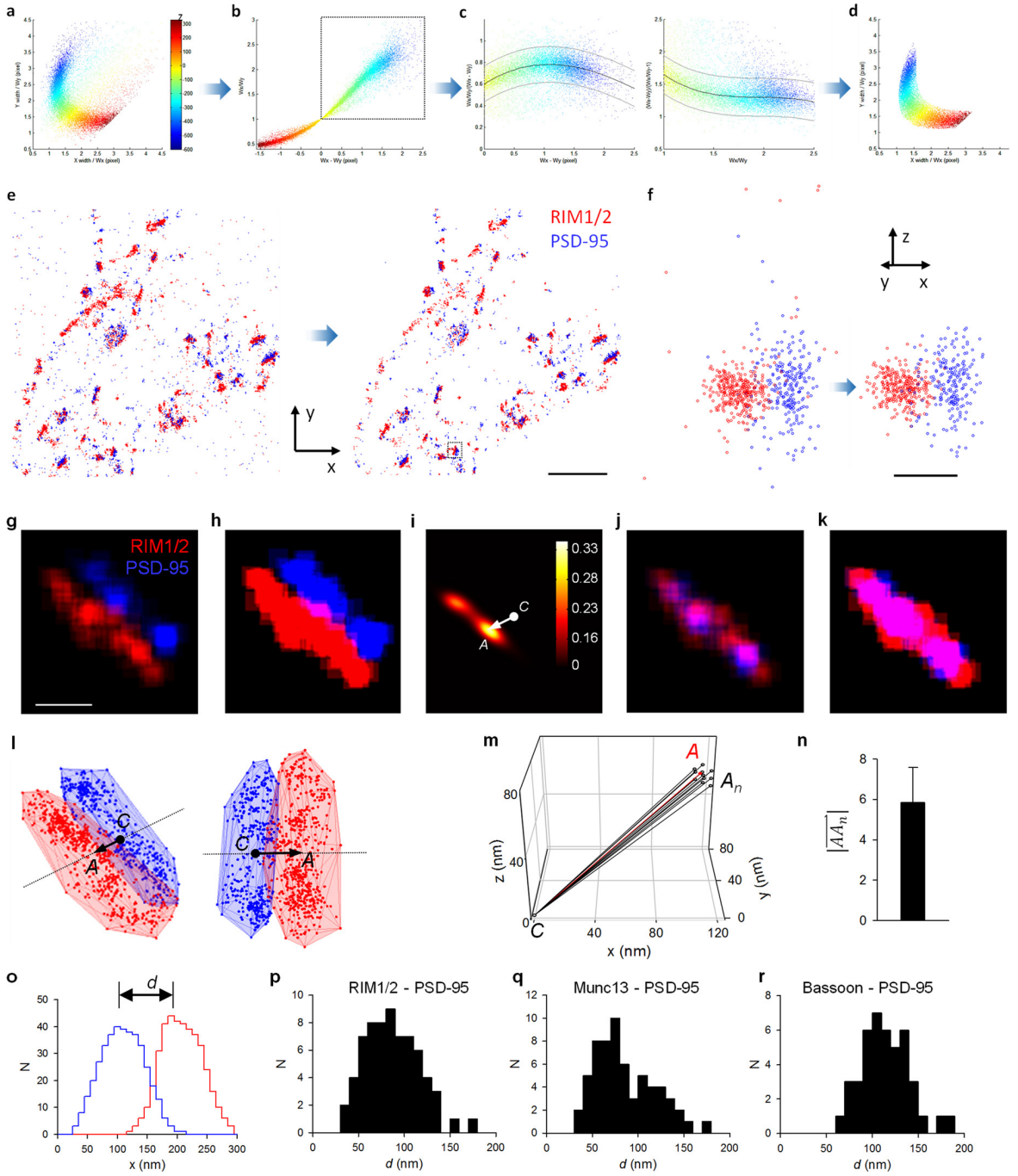
where s is the steepness of the sigmoid transition, LD_i is the normalized local density of RIM at the ith pixel of the presynaptic gradient map, $LD_{inflect}$ is the point of inflection in the sigmoidal function, and LD_{max} is the maximum normalized local density of RIM in the STORM measured example shown in Extended Data Fig. 8b. $LD_{inflect}$ and s were fitted to be 1.5 and 0.959 in order to yield a fractional release area of 40%. To calculate the average peak synaptic response per release, we calculated the expected open channel fraction averaged over all possible release sites weighted by the spatial probabilities of release:

$$\text{Open channels at peak response } (\%) = \sum_i \left[O(i) \frac{P_r(i \mid \text{release})}{\sum_i P_r(i \mid \text{release})} \right]$$

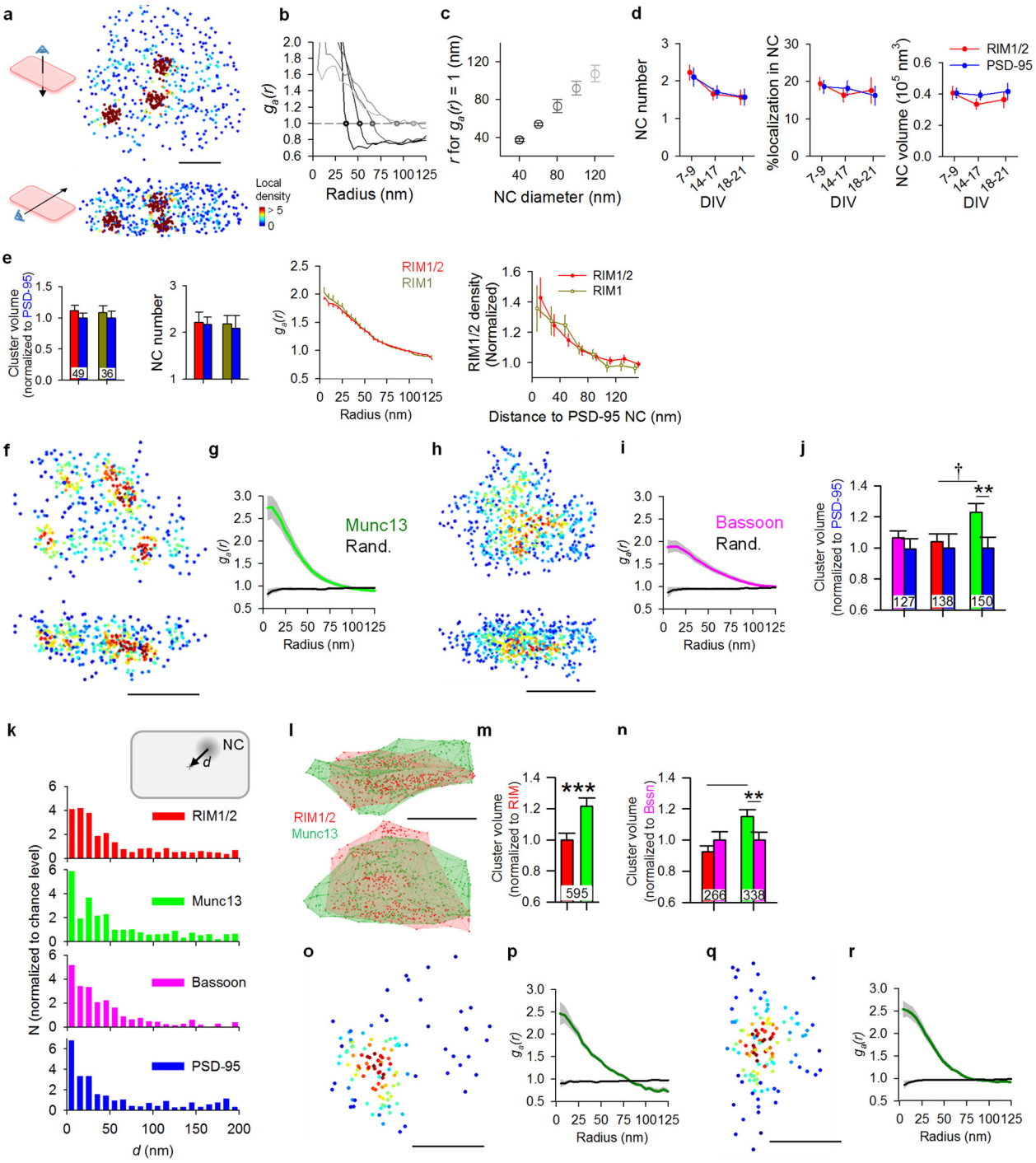
Code availability. All code used in the paper is available upon request.

Statistical analysis. No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment. Statistical tests were performed with Sigmastat, MATLAB, Graphpad, or R. No statistical methods were used to predetermine sample size. The sample sizes were determined based on numbers reported in previous studies. For comparison of two or more distributions, all samples were assessed for normality using Shapiro–Wilk or Kolmogorov–Smirnov tests. If samples met criteria for normality, we used a Student's t-test to compare two groups, a paired t-test for comparison of the same group before and after a treatment, or ANOVA for more than two groups. If ANOVAs were significant, we used a post hoc Tukey test to compare between groups. For groups with combinations of discrete and continuous variables, we used MANCOVAs. We only performed two-tailed tests. Homogeneity of variances was tested using an F-test and found to be similar between compared groups. If samples did not meet criteria for parametric tests, we used Kolmogorov–Smirnov or Wilcoxon rank-sum tests for comparison of two groups and Kruskal–Wallis or Friedman ANOVA for comparison of more than two groups, with post hoc analysis using Dunn's test. Data are presented as mean ± s.e.m. unless otherwise specified. Also see Supplementary Tables.

31. Frost, N. A., Shroff, H., Kong, H., Betzig, E. & Blanpied, T. A. Single-molecule discrimination of discrete perisynaptic and distributed sites of actin filament assembly within dendritic spines. *Neuron* **67,** 86–99 (2010).
32. Kaech, S. & Banker, G. Culturing hippocampal neurons. *Nat. Protocols* **1,** 2406–2415 (2006).
33. van de Linde, S. *et al.* Direct stochastic optical reconstruction microscopy with standard fluorescent probes. *Nat. Protocols* **6,** 991–1009 (2011).
34. Schneider Gasser, E. M. *et al.* Immunofluorescence in brain sections: simultaneous detection of presynaptic and postsynaptic proteins in identified neurons. *Nat. Protocols* **1,** 1887–1897 (2006).
35. Geisler, C. *et al.* Drift estimation for single marker switching based imaging schemes. *Opt. Express* **20,** 7274–7289 (2012).
36. Mlodzianoski, M. J. *et al.* Sample drift correction in 3D fluorescence photoactivation localization microscopy. *Opt. Express* **19,** 15009–15019 (2011).
37. Thompson, R. E., Larson, D. R. & Webb, W. W. Precise nanometer localization analysis for individual fluorescent probes. *Biophys. J.* **82,** 2775–2783 (2002).
38. Schikorski, T. & Stevens, C. F. Quantitative ultrastructural analysis of hippocampal excitatory synapses. *J. Neurosci.* **17,** 5858–5867 (1997).
39. Veatch, S. L. *et al.* Correlation functions quantify super-resolution images and estimate apparent clustering due to over-counting. *PLoS One* **7,** e31457 (2012).
40. Kim, S. H. & Ryan, T. A. CDK5 serves as a major control point in neurotransmitter release. *Neuron* **67,** 797–809 (2010).
41. Voglmaier, S. M. *et al.* Distinct endocytic pathways control the rate and extent of synaptic vesicle protein recycling. *Neuron* **51,** 71–84 (2006).
42. Harris, K. M. & Stevens, J. K. Dendritic spines of CA 1 pyramidal cells in the rat hippocampus: serial electron microscopy with reference to their biophysical characteristics. *J. Neurosci.* **9,** 2982–2997 (1989).
43. Simonson, P. D., Rothenberg, E. & Selvin, P. R. Single-molecule-based super-resolution images in the presence of multiple fluorophores. *Nano Lett.* **11,** 5090–5096 (2011).
44. Thompson, M. A., Lew, M. D. & Moerner, W. E. Extending microscopic resolution with single-molecule imaging and active control. *Annu. Rev. Biophys.* **41,** 321–342 (2012).
45. Chen, T.-W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499,** 295–300 (2013).
46. Murthy, V. N., Sejnowski, T. J. & Stevens, C. F. Dynamics of dendritic calcium transients evoked by quantal release at excitatory hippocampal synapses. *Proc. Natl Acad. Sci. USA* **97,** 901–906 (2000).
47. Xie, X., Liaw, J.-S., Baudry, M. & Berger, T. W. Novel expression mechanism for synaptic potentiation: alignment of presynaptic release site and postsynaptic receptor. *Proc. Natl Acad. Sci. USA* **94,** 6983–6988 (1997).
48. Kavalali, E. T. *et al.* Spontaneous neurotransmission: an independent pathway for neuronal signaling? *Physiology (Bethesda)* **26,** 45–53 (2011).
49. Frank, T. *et al.* Bassoon and the synaptic ribbon organize Ca^{2+} channels and vesicles to add release sites and promote refilling. *Neuron* **68,** 724–738 (2010).
50. Ermolyuk, Y. S. *et al.* Differential triggering of spontaneous glutamate release by P/Q-, N- and R-type Ca^{2+} channels. *Nat. Neurosci.* **16,** 1754–1763 (2013).

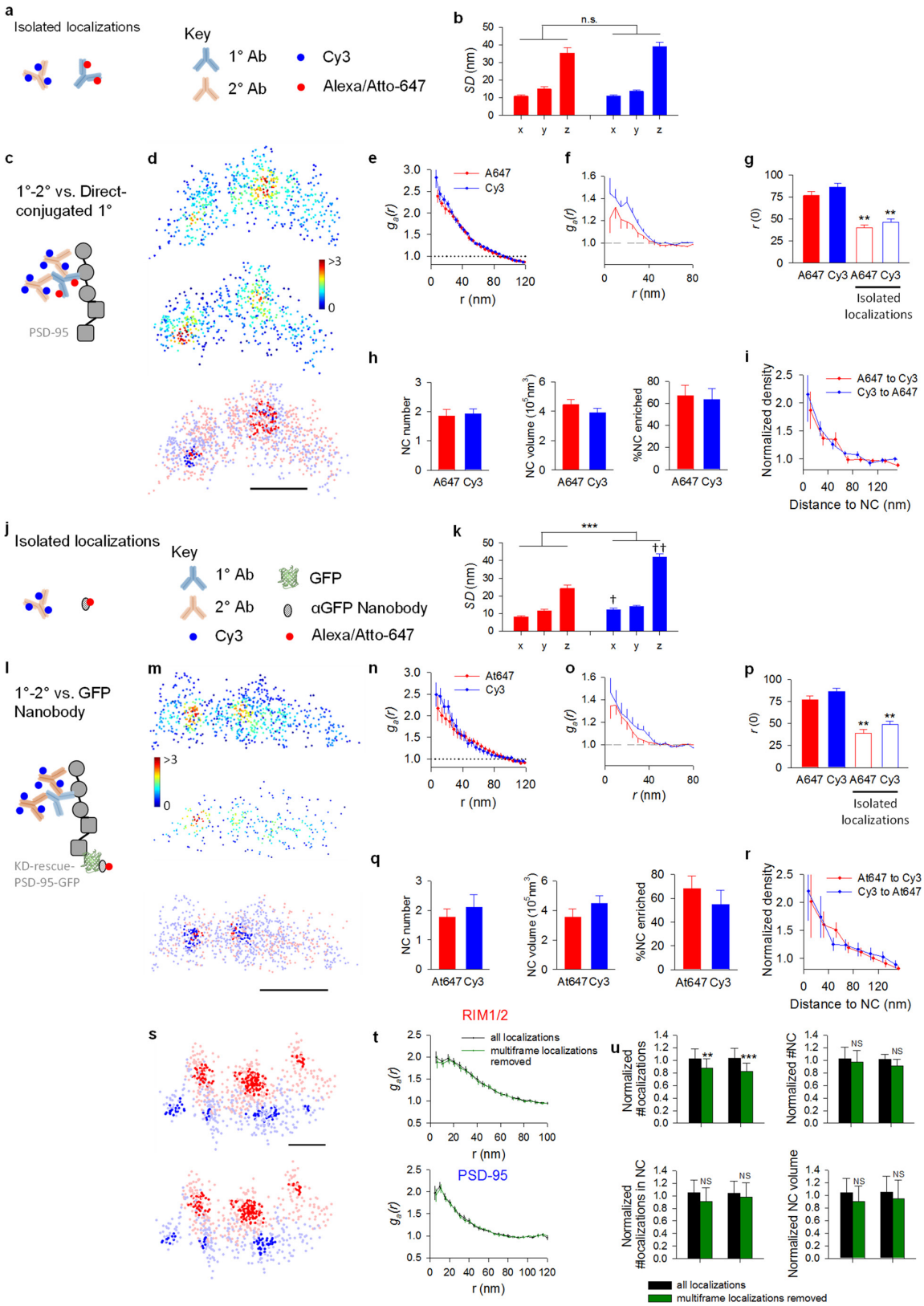


Extended Data Figure 1 | Filtering of localizations and automatic algorithm to detect the synaptic axis. **a**, Scatter plot of fitted peak width in y (W_y) against that in x (W_x). The colour codes the position in z. All localizations away from this centre dense region arise from multiple overlapping or poorly fitted peaks and should be rejected. **b**, The ellipticity (W_x/W_y) and the width difference ($W_x - W_y$) formed an approximate linear relationship when $W_x > W_y$ (dotted box). **c**, We fitted the ratios between ellipticity and the width difference to the denominators with third degree polynomial functions (black line) and rejected all localizations out of 95% confidence intervals (grey lines) of the curve (>1.96 × s.d.). The same criteria was applied to the other fraction of localizations with $W_x < W_y$. **d**, The same scatter plot as in **a** after rejection of all of the diffuse localizations (about 20–25%). **e**, **f**, The filtering protocol cleared up most of the localizations from multiple overlapping peaks or poorly fitted peaks, including most of the non-relevant background localizations (**e**) and those localizations with poorly calibrated z positions (**f**). Scale bars, 2 μm (**e**) and 200 nm (**f**). The synapse in **f** corresponds to the boxed synapse in **e**. **g**, A 2D section through the centre of the convoluted constructed 3D distribution matrix of a synapse. **h**, Peak density of the matrix set to a quarter of the mean molecule density of the synaptic cluster. **i**, 2D section at the same position of the 3D matrix of direct cross-correlation of the two channels (equation (3) in Methods). C is the centre of matrix, and A is the peak of the cross-correlation. **j**–**k**, Best overlap of the two synaptic clusters after PSD-95 was moved in 3D space along the vector $\overrightarrow{CA}$. **l**, 3D scatter plots of the synapse in two different view angles. The arrow denotes the vector and the extended line (dotted) represents the synaptic axis. **m**, 3D plot of detected synaptic axis when the positions of high-density peaks in RIM1/2 (nanoclusters) were randomized within the synaptic cluster. This simulation was performed 35 times, but only 10 representative results are presented here to avoid overlapping. The red denotes the synaptic axis of the original synaptic cluster. **n**, Averaged distance between the detected C_n positions from 35 simulated clusters to the C position of the original cluster. Data shown in mean ± s.d. This <6 nm distance confirms that the high-density peaks have negligible effect on the detection of the synaptic axis in this Method. **o**, Distribution of all localizations along the synaptic axis with bin size of 10 nm. Peak-to-peak distance between the synaptic protein pair can be measured from this distribution. **p**–**r**, Distribution of peak-to-peak distances for three pairs of synaptic proteins.



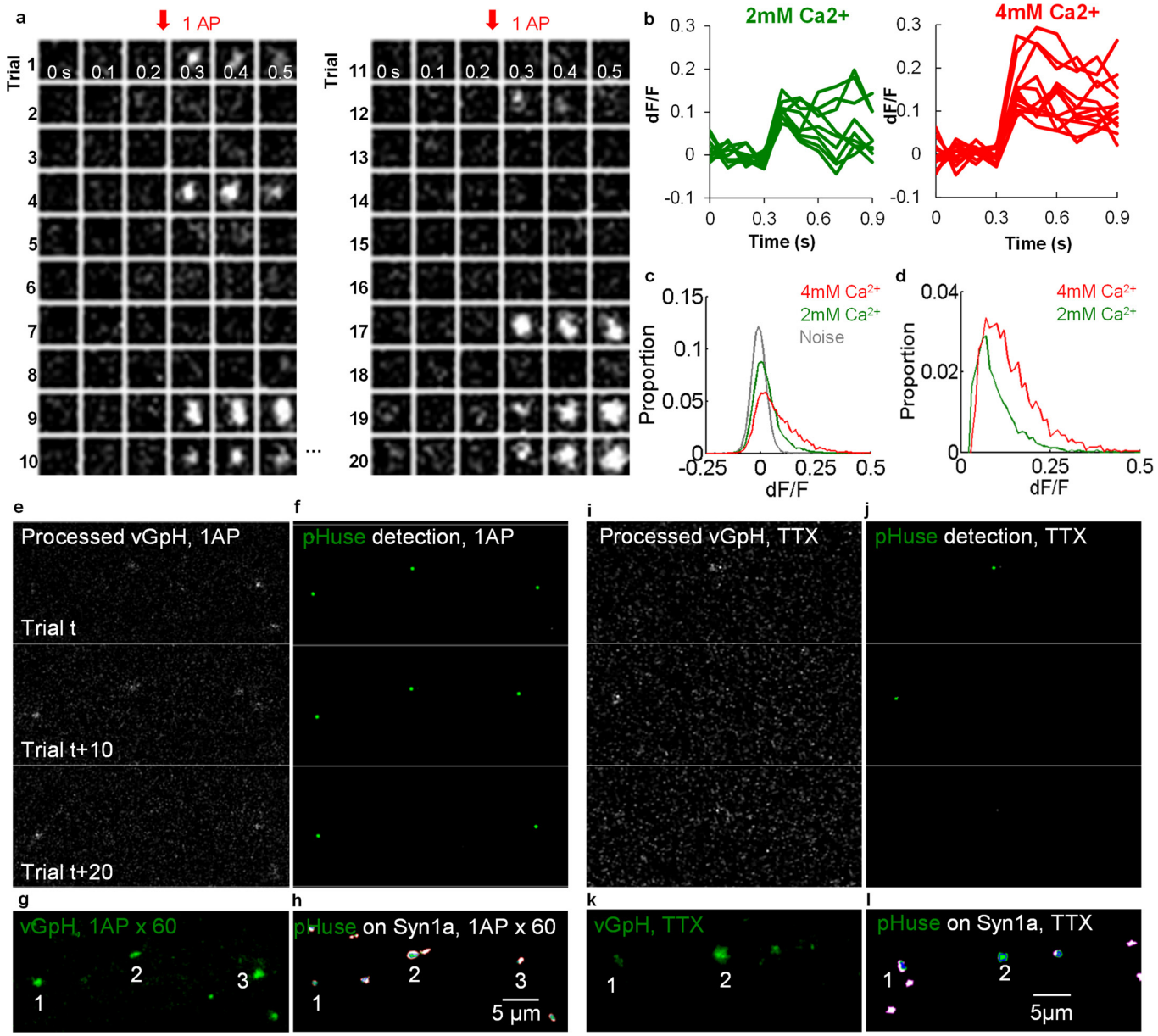
Extended Data Figure 2 | See next page for caption.

Extended Data Figure 2 | Nanocluster organization of vesicle release machinery proteins in the active zone and postsynaptic AMPA receptors. **a**, *En face* (top) and side (bottom) views of local density maps of a simulated synapse with artificial nanoclusters with 40-nm diameters. Scale bar, 100 nm. **b**, Autocorrelation function of simulated clusters with different sized nanoclusters. The points represent the radius where $g(r) = 1$. **c**, Pooled data from 15 sets of simulations showing that the radius where $g(r)$ first crosses 1 reasonably estimates the average nanocluster diameters. **d**, Comparison of nanocluster number, fraction of localization in nanocluster, and nanocluster volume across different developmental stages shows no significant difference, though the young 9 days *in vitro* (DIV) culture shows a trend towards increased nanocluster numbers (one-way ANOVA on ranks for nanocluster number and volume, one-way ANOVA for percentage localization in nanocluster). Data were from 143 RIM nanoclusters and 135 PSD nanoclusters of 64 DIV 9 synapses, 63 RIM nanoclusters and 65 PSD nanoclusters of 38 DIV 14 synapses, and 44 RIM nanoclusters and 41 PSD nanoclusters from 28 DIV 21 synapses. **e**, Comparison of two RIM antibodies (from left to right) in whole synaptic cluster volume, number of nanoclusters, autocorrelation function estimating average nanocluster diameter, and protein density relative to PSD-95 nanocluster centres. Anti-RIM1/2 (Synaptic Systems #140-203) targets the zinc-finger domain and anti-RIM1 targets the PDZ domain of RIM1 (Synaptic Systems #140-003). These tests suggest that there is no significant difference between these two antibodies. The numbers in bars denote the group sizes. **f**, Local density maps of *en face* (top) and side (bottom) views of an example Munc13 cluster. Scale bar, 200 nm. **g**, Auto-correlation functions for Munc13 distributions compared to simulated randomized distributions. **h**, **i**, Local density maps and ACF of Bsn cluster. Scale bar, 200 nm. **j**, Pooled cluster volumes, normalized to PSD-95 volumes within each synapse. Each bar pair represents data from a set of RIM1/2-PSD-95, Munc13-PSD-95 or Bsn-PSD-95 staining. The numbers in bars denote the group sizes. **k**, Distribution of *en face* distances between nanocluster centre and synapse centre. Data were normalized to the distribution of simulated clusters with the same number of nanoclusters as the original synapse but randomized positions. **l**, An example synapse with RIM1/2 and Munc13 staining of the same synapse, shown in two different angles. The translucent surfaces represent the alpha shapes that define the synaptic cluster borders. **m**, Pooled RIM1/2 and Munc13 cluster volumes, normalized to RIM1/2 within each synapse. **n**, Pooled RIM1/2, Munc13 and Bsn cluster volumes from staining of RIM1/2-Bsn and Munc13-Bsn, normalized to Bsn within each synapse. $^{*}P < 0.05$; $^{***}P < 0.001$; Wilcoxon signed-rank test. $\dagger P < 0.05$, one-way ANOVA on ranks with pairwise comparison procedures (Dunn's method). **o**, Local density map of a GluA2 cluster. **p**, Auto-correlation functions for GluA2 distributions compared to simulated randomized distributions. **q**, Local density map of a GluR2/3 cluster. **r**, Auto-correlation functions for GluR2/3 distributions compared to simulated randomized distributions. All experiments were repeated $\geq$3 times.

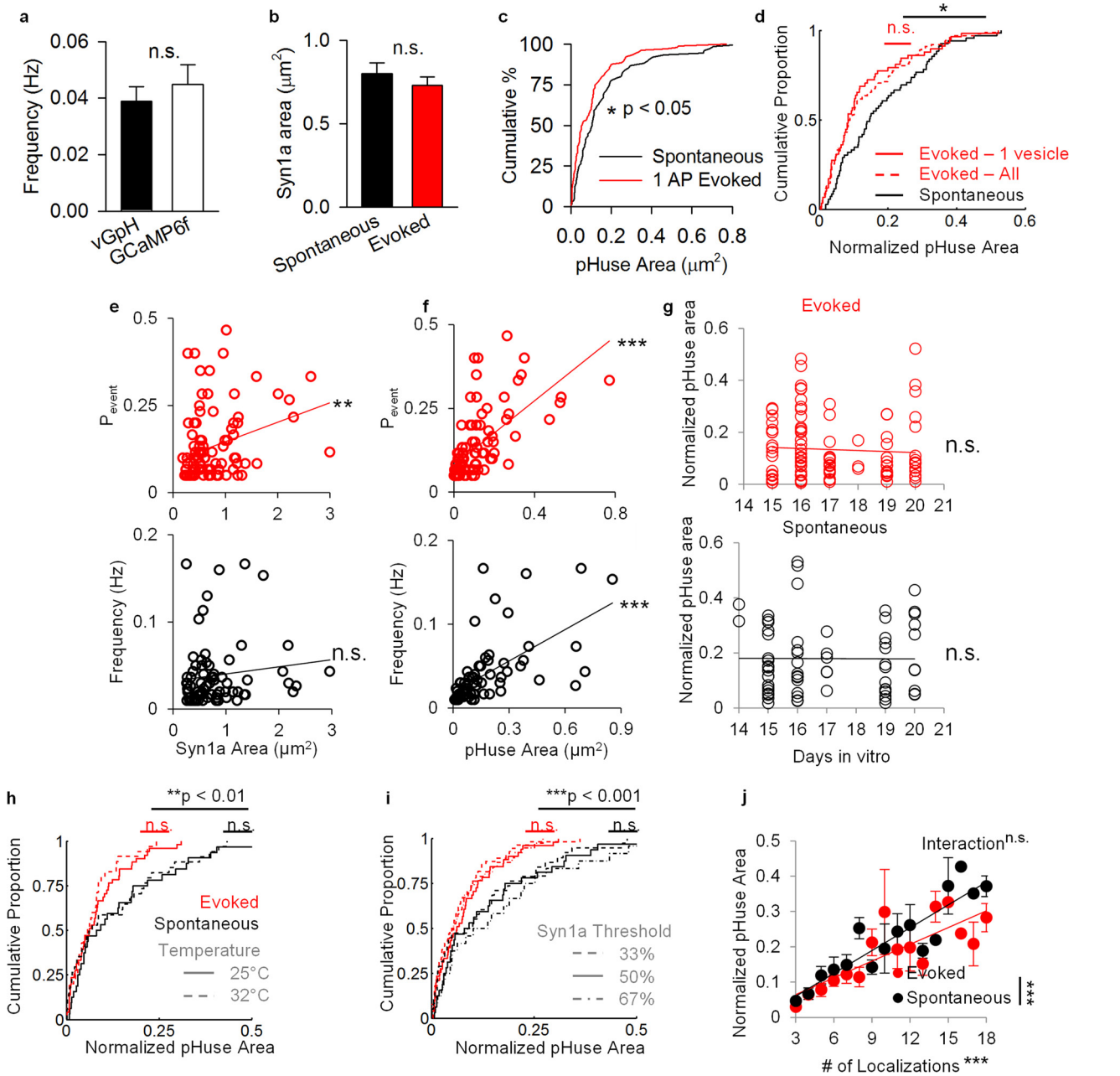


Extended Data Figure 3 | See next page for caption.

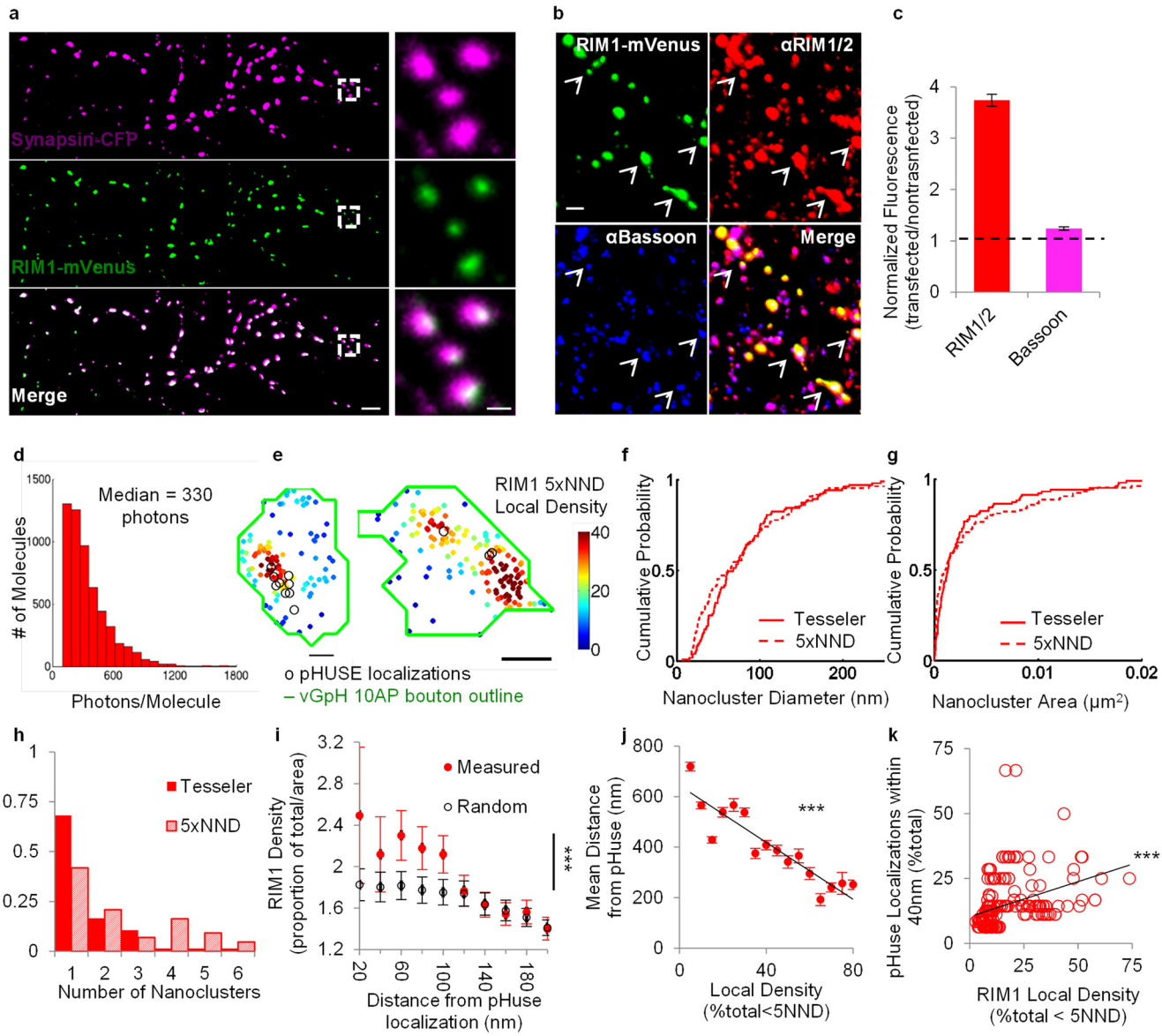
Extended Data Figure 3 | Detected nanoclusters are unlikely a result of labelling artefacts or overcounting of molecules. a–i, Comparison of PSD-95 labelled with monoclonal primary antibodies directly conjugated to Alexa647 dye (1°-A647, red) with the same molecules labelled with primary and secondary antibodies conjugated to Cy3 (1°-2°-Cy3, blue) as represented in **c**. **a**, **b**, Comparison between non-synaptic small groups of localizations arising from isolated primary antibodies and secondary antibodies. Schematic shown in **a**. Standard deviation of localizations in both groups along different dimensions ($n = 32$ for A647; $n = 36$ for Cy3) in **b**. The two types of localizations groups showed similar variation in all dimensions. **d**, Local density maps of the same PSD-95 cluster labelled with 1°-A647 (top) and 1°-2°-Cy3 (middle) and overlapped distribution of 1°-A647 and 2°-Cy3 with detected nanoclusters highlighted in darker colours (bottom). Scale bar, 200 nm. **e**, Autocorrelation of synaptic clusters labelled with 1°-A647 and 1°-2°-Cy3. **f**, Autocorrelation of isolated small groups of localizations of A647 and Cy3 dyes. **g**, Comparison of the radius at which the autocorrelation function crossed with the random level ($g(r) = 1$). There was no difference between PSD-95 clusters with different labelling methods, but the $r(0)$ for isolated localization groups were significantly less than $r(0)$ for PSD-95 clusters. $^{**}P < 0.01$, *t*-test between the filled and open bars of the same colour. **h**, Nanoclusters detected in both channels displayed no difference in number, volume, or the fraction of nanoclusters enriched with localizations from the other channel. **i**, Protein enrichment of localizations detected in each channels with those in the other channel ($n = 32$ synapses). These results demonstrate that the nanoclusters we detected in our study were not due to aggregation of multiple secondary antibodies to the primary antibodies. **j–r**, Cells transfected with knockdown-rescue-PSD-95-GFP were labelled with nanobodies against GFP conjugated at a 1:1 ratio with Atto647 (Nb-At647, red) and primary/secondary antibodies against PSD-95 (1°-2°-Cy3, blue) as depicted in **l**. **j**, **k**, Comparison between non-synaptic small groups of localizations arising from isolated Nb-At647 and 1°-2°-Cy3 (as depicted in **j**, $n = 26$ and 28, respectively). **k**, The nanobodies showed a significant smaller size than antibodies. $^{***}P < 0.001$, two-way ANOVA, $^{\dagger}P < 0.05$, $^{\dagger\dagger}P < 0.01$, pairwise comparison (Tukey test) between nanobodies and antibodies. **m–r**, Similar comparison as in **d–i** between PSD-95 clusters labelled with Nb-At647 and 1°-2°-Cy3 ($n = 13$ synapses). Scale bar, 200 nm. Overall, these results demonstrated that the nanoclusters we detected in our study were unlikely a result of artefacts of antibody binding and labelling. The difference between the size of the isolated localizations groups and PSD-95 clusters calculated by autocorrelation also argues against the possibility that the nanoclusters we detected were owing to repetitive switching of one or a few fluorophores. $^{**}P < 0.01$, *t*-test between the filled and open bars of the same colour. **s**, An example synapse with nanoclusters highlighted before (upper) and after (lower) removal of localizations resulting from fluorophores lasting for multiple frames. Scale bar, 100 nm. **t**, Paired autocorrelation function of synaptic clusters with and without multiple-frame molecules. $P = 0.77$, $n = 25$ synapses for RIM1/2; $P = 0.58$, $n = 25$ synapses for PSD-95, two-way ANOVA with repeated measures. **u**, The tracking removed $13 \pm 8\%$ and $17 \pm 9\%$ of the localizations for RIM1/2 and PSD-95, respectively, but had no significant effects on autocorrelation function results, nanocluster numbers, or nanocluster volumes. $^{**}P < 0.01$; $^{***}P < 0.001$; NS, $P > 0.05$; Wilcoxon signed-rank test. All data were pooled from ≥ 3 replicas.



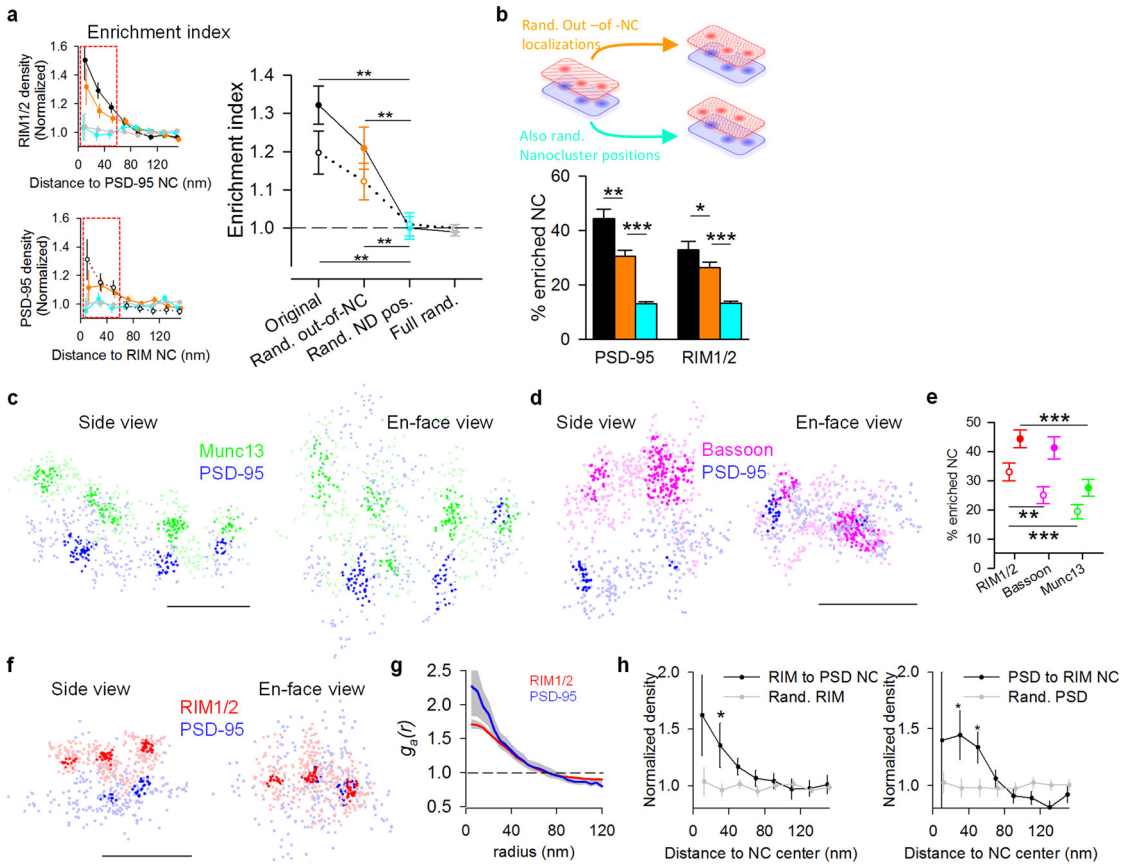
Extended Data Figure 4 | 1AP evoked release is $[Ca^{2+}]$ dependent and mainly univesicular[48]. **a**, Example of fluorescence signals at a single bouton over repeated trials of 1 action potential stimulation. **b**, Single event traces of vGpH fluorescence increase following 1 action potential stimuli in standard (2 mM) or heightened extracellular $[Ca^{2+}]$ (4 mM). **c**, Comparison of distributions of fluorescence changes in 2 mM ($n = 233/27$) and 4 mM ($n = 115/12$) extracellular $[Ca^{2+}]$, relative to noise distributions obtained from the baseline frames before stimulation. **d**, Comparison of noise-subtracted distributions of fluorescence changes in different $[Ca^{2+}]$. **e**, Processed images of vGpH fluorescence increase following 1 action potential stimuli from three trials ten trials apart. **f**, Automatic detection using pHuse of events shown in **e**. **g**, Summed projection of framewise and background subtracted vGpH fluorescence increases over 60 trials. **h**, pHuse localizations on Syn1a (white). **i–l**, Same as **e–h** for spontaneous events in TTX over 5 min. *n* given in synapses/experiments.



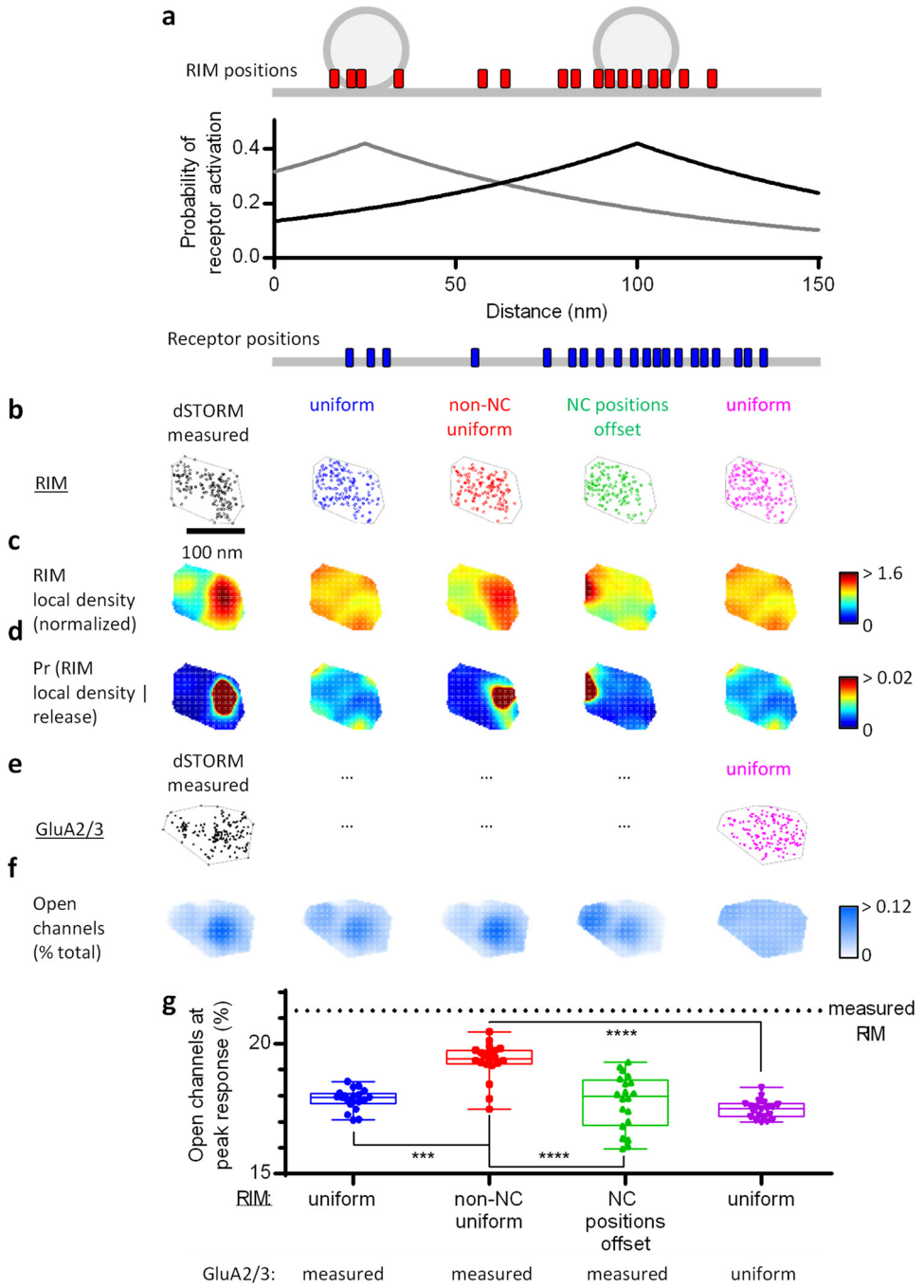
Extended Data Figure 5 | pHuse reveals differences between evoked and spontaneous fusion site areas. a, Comparison of spontaneous frequency measured presynaptically using vGpH ($n = 77/22$) and postsynaptically using GCaMP6f (ref. 45) ($n = 61/5$), $t = 1.02$, not significant. **b**, Average bouton areas across groups, $t = 0.87$, not significant. **c**, Cumulative distributions of fusion areas for spontaneous and evoked release (Kolmogorov–Smirnov test, $^*D = 0.23$) **d**, Cumulative distributions of normalized fusion areas for 1 AP evoked fusion excluding events with photon counts > mean + 2 s.d. of spontaneous events ($n = 91/27$) compared to all evoked events ($n = 104/28$, Kolmogorov–Smirnov test, $D = 0.05$, not significant) and spontaneous events ($n = 77/22$, Kolmogorov–Smirnov test, $^*D = 0.25$) **e**, **f**, Notably, while evoked P_r was significantly positively correlated with Syn1a area, as reported previously[49], spontaneous event frequency showed no relationship with Syn1a area (**e**, linear fit, evoked $^{**}R = 0.30$, spontaneous $R = 0.12$, not significant). On the other hand, both spontaneous event frequency and evoked P_r significantly positively correlated with pHuse area (**f**, linear fit, evoked $^{***}R = 0.64$, spontaneous $^{***}R = 0.60$). This suggests that pHuse area may be a better approximation for active zone area and the functional parameters of a synapse than bouton area. **g**, Normalized pHuse area as a function of cell age shows no significant correlation (evoked $R = 0.03$, not significant, spontaneous $R = 0.004$, not significant). **e**–**g**, $n_{evoked} = 104/28$, $n_{spont} = 77/22$. **h**. Normalized pHuse area was not significantly different at room temperature ($n_{evoked} = 51/10$, $n_{spont} = 32/7$) versus physiological temperature ($n_{evoked} = 35/9$, $n_{spont} = 34/4$) within modes of release but still significantly different between modes of release. **i**, Normalized pHuse area was not significantly different at different thresholds for Syn1a within modes of release but still significantly different between modes of release ($n = 51/10$). **j**, Both numbers of events and mode of release are significant factors for pHuse area, but they do not have a significant interaction $n_{evoked} = 155/38$, $n_{spont} = 109/29$. For **i**, **j**, see Supplementary Tables for statistics. n given in synapses/experiments, $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.



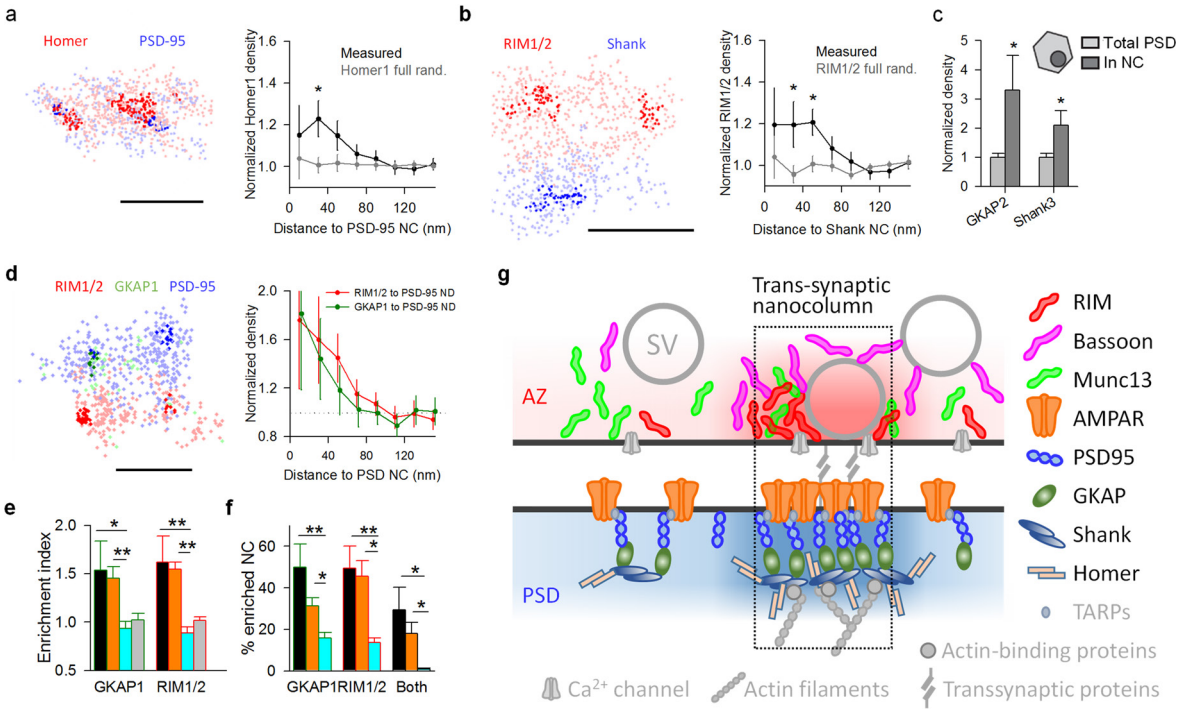
Extended Data Figure 6 | RIM1-mEos3.1 PALM identifies nanoclusters. **a**, Neurons co-expressing RIM1-mVenus (a gift from P. Kaesar) and Syn1a–CFP colocalize to the same boutons. Right panels show enlargement of areas within the white boxes. Scale bars, 5 μm (left) and 1 μm (right). **b**. Neurons expressing RIM1-mVenus immunostained for RIM1/2 and Bsn. Arrowheads point to some colocalized active zones. Scale bar, 2 μm. **c**, Immunofluorescence intensity of transfected cells normalized to nearby untransfected cells show 3.74 ± 0.11-fold overexpression of RIM and 1.24 ± 0.03-fold increase in Bsn ($n = 262$ synapses/7 cells). **d**, Photon count distribution of RIM1-mEos3.1 (3997 localizations). **e**, Same boutons shown in Fig. 2 visualized using 5 × nearest neighbour density (NND) as a measure of local density. **f**–**h**, Cumulative distributions of PALMed RIM1 nanoclusters diameter, area, and number, respectfully, identified using adapted Tesseler analysis and 5 × NND analysis ($n = 65/13$). **i**, RIM1 localization density as a function of radial distance from pHuse localizations. (See Supplementary Tables for statistics.) **j**, Mean distance from pHuse localizations as a function of local density measured by 5 × NND (raw data ***$R = 0.23$, $n = 26/13$). **k**, Proportion of pHuse localizations within 40 nm of a RIM1 localization as a function of RIM1 local density measured by 5 × NND (***$R = 0.35$). *n* given in synapses/experiments unless otherwise specified, ***$P < 0.001$.



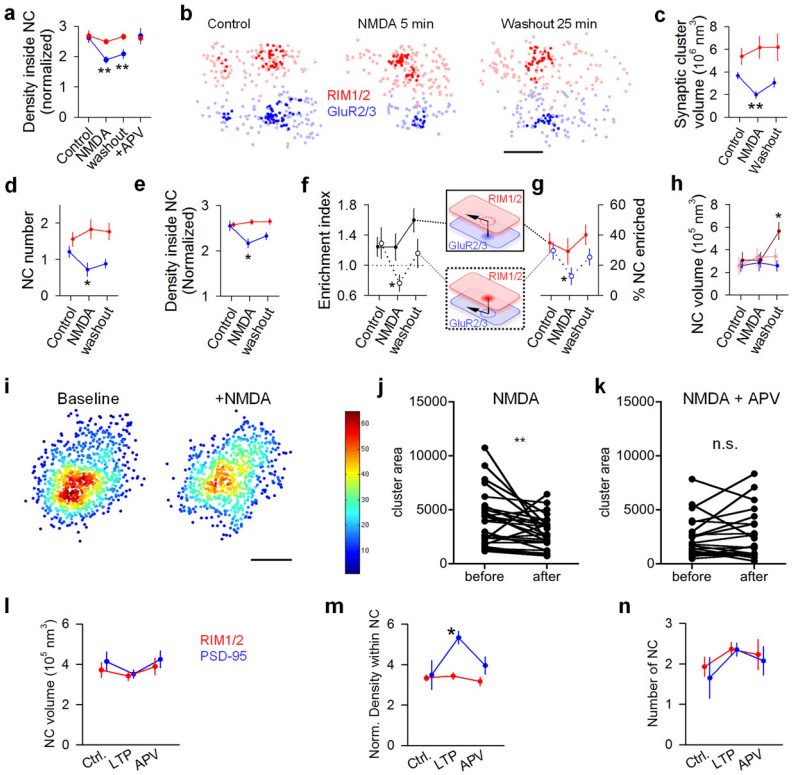
Extended Data Figure 7 | Protein enrichment within nanocolumns. **a**, Enrichment index between RIM1/2 and PSD-95. The left insets are replicas of Fig. 3e, and the enrichment index is defined as the average of the first three bins in the enrichment profile (boxed), that is, normalized localization density within 60 nm from the projection centre of a given nanocluster. Filled points show RIM1/2 relative to PSD-95 nanoclusters, open points show PSD-95 relative to RIM1/2 nanoclusters. Same randomizations as in Fig. 3e and depicted again in **b**. $^{**}P < 0.01$; $^{***}P < 0.001$, one-way ANOVA on ranks with pairwise comparison procedures (Dunn's method). **b**, The fraction of enriched nanoclusters is significantly above chance level, and is also dependent on the relative position of the two sets of nanoclusters. **c**, **d**, Side and *en face* views of a synaptic Munc13 and PSD-95 pair and a synaptic Bsn and PSD-95 pair with highlighted nanoclusters. Scale bar, 200 nm. **e**, Pooled enrichment index of three active zone proteins and PSD-95. Scale bar, 200 nm. Filled points show active zone proteins relative to PSD-95 nanoclusters, open points show PSD-95 relative to active zone protein nanoclusters. $^{**}P < 0.01$; $^{***}P < 0.001$, one-way ANOVA on ranks with pairwise comparison procedures (Dunn's method). **f**, Example of RIM1/2 and PSD-95 in adult hippocampal slices. **g**, Auto-correlation functions of RIM1/2 and PSD-95 ($n = 192$ and 43 synapses, respectively). There were, on average, 2.02 ± 0.08 and 1.32 ± 0.21 nanoclusters with a volume of (3.6 ± 0.2) and $(4.2 \pm 0.7) \times 10^5$ nm^3 for RIM1/2 and PSD-95, respectively. Except PSD nanocluster number which was significantly less than that in cultures ($P = 0.03$), all other parameters were similar (Wilcoxon signed-rank test). **h**, Enrichment profile between RIM1/2 and PSD-95 in tissue slices (28 synapses from 7 sections, 4 animals). $^{*}P < 0.05$ between measured and randomized synapses, two way ANOVA with pairwise comparison procedures (Dunn's method).



Extended Data Figure 8 | Preferential release in nanocolumns can increase synaptic strength. a, Schematic of the experimentally constrained, deterministic approach used to study the dependence of synaptic strength on the spatial distribution of release sites and AMPARs. The simulated release site distribution at a synapse was drawn from its measured RIM positions and the average measured relationship between RIM density and pHuse locations (Fig. 2). **b**, Distributions of measured RIM localizations within a single active zone (active zone) boundary (grey), and the same cluster with randomized positions of the indicated subsets of molecules. **c**, Maps of RIM local density normalized to the overall densities within the active zones. **d**, Probability density maps of possible release sites given that a release occurs. **e**, Distributions of GluA2/3 locations within the PSD boundary (grey) of the same measured synapse (ellipses refer to this distribution) and randomized. **f**, Maps of fraction of open channels at peak response per average release from the respective active zones directly above them in **d**. **g**, Calculated open channels at peak response, $n = 20$ randomly generated molecular distributions. See Methods for more details.



Extended Data Figure 9 | Enrichment of other scaffolding proteins within nanocolumns. a, Enrichment of Homer1 with PSD-95 nanoclusters, $n = 118$ nanoclusters from 48 synapses, scale 100 nm. **b**, Enrichment of RIM1/2 to Shank nanoclusters, $n = 80$ nanoclusters from 32 synapses. Scale bar, 200 nm. $^*P < 0.05$, ANOVA on ranks with pairwise comparison procedures (Dunn's method) in **a** and **b**. **c**, GKAP2 and Shank3 densities (determined with STORM, $n = 6$ and 12, respectively) within PSD-95 nanoclusters (determined with PALM of transfected knockdown-replacement-PSD-95-mEos2) normalized to total PSD densities. Both proteins showed significant enrichment in PSD-95 nanoclusters, $^*P < 0.05$, paired t-tests. **d**, Three-colour STORM imaging of RIM1/2, GKAP1 and PSD-95 on the same synapses example (left) and protein enrichment profiles of RIM1/2 and GKAP1 with respect to PSD-95 nanoclusters (right), $n = 32$ nanoclusters from 17 synapses. Scale bar, 200 nm. **e**, Enrichment indices of RIM1/2 and GKAP1 relative to PSD-95 nanoclusters. Colour-coded bars represent the same set of randomizations as performed in Fig. 3c: orange denotes randomization of only out-of-nanocluster localizations, cyan denotes randomization of nanocluster positions within synaptic clusters and grey denotes randomization of all localizations. **f**, The percentage of PSD-95 nanoclusters that were enriched with GKAP1, RIM1/2 or both with colour-coded randomizations. $^*P < 0.05$; $^{**}P < 0.01$, ANOVA on ranks with pairwise comparison procedures (Dunn's method), $n = 32$ nanoclusters from 17 synapses in 7 different cultures. **g**, Schematic summary of the distribution of synaptic proteins within nanocolumns. The distributions of colour-coded proteins are based on our results and the proteins in grey are hypothetical, some, such as Ca^{2+} channels, have been suggested previously to be clustered[49,50]. All experiments were repeated ≥ 5 times.



Extended Data Figure 10 | Plasticity within nanocolumns. a, Changes in the localization density within RIM1/2 (red) and PSD-95 (blue) nanoclusters under control, 5 min NMDA treatment, 25 min washout, and NMDA + AP5 treatment conditions. **b**–**h**, Reorganization of RIM1/2 and GluR2/3 under control, 5 min NMDA treatment, 25 min washout conditions examples (**b**), comparison of whole synaptic cluster sizes (**c**), nanocluster number per synapse (**d**), localization density within nanoclusters (**e**), enrichment indices (**f**), percentage of nanoclusters that were enriched (**g**), and nanocluster volumes (**h**). Note that similar to the results from the RIM1/2-PSD-95 analyses, only those RIM1/2 nanoclusters that were enriched with GluR2/3 (dark red) were increased in volume. $^{*}P < 0.05$; $^{**}P < 0.01$, ANOVA on ranks with pairwise comparison to control group (Dunn's method), and χ^2 test for the proportion. Data from 62, 21 and 37 nanoclusters from 34, 18 and 24 synapses for control, NMDA, and washout, respectively. **i**, Colour-coded local density map of an example live-PALMed PSD-95 cluster before and after NMDA treatment. Scale bar, 100 nm. **j**, **k**, Changes in PSD-95 nanocluster area induced by NMDA and blocked by AP5 ($n = 28$ and 21, respectively). $^{**}P < 0.01$, NS, not significant, paired t-test. **l**–**n**, LTP stimulation induced changes in nanocluster volumes (**l**), localization density within nanoclusters (**m**) and nanocluster numbers (**n**). $^{*}P < 0.05$, ANOVA on ranks with pairwise comparison to control group (Dunn's method). All experiments were repeated $\geq$5 times.

LETTER

doi:10.1038/nature19076

Tumour-cell-induced endothelial cell necroptosis via death receptor 6 promotes metastasis

Boris Strilic[1], Lida Yang[1], Julián Albarrán-Juárez[1], Laurens Wachsmuth[2], Kang Han[3], Ulrike C. Müller[3], Manolis Pasparakis[2] & Stefan Offermanns[1,4]

Metastasis is the leading cause of cancer-related death in humans. It is a complex multistep process during which individual tumour cells spread primarily through the circulatory system to colonize distant organs[1–3]. Once in the circulation, tumour cells remain vulnerable, and their metastatic potential largely depends on a rapid and efficient way to escape from the blood stream by passing the endothelial barrier[4–9]. Evidence has been provided that tumour cell extravasation resembles leukocyte transendothelial migration[7–9]. However, it remains unclear how tumour cells interact with endothelial cells during extravasation and how these processes are regulated on a molecular level. Here we show that human and murine tumour cells induce programmed necrosis (necroptosis) of endothelial cells, which promotes tumour cell extravasation and metastasis. Treatment of mice with the receptor-interacting serine/threonine-protein kinase 1 (RIPK1)-inhibitor necrostatin-1 or endothelial-cell-specific deletion of RIPK3 reduced tumour-cell-induced endothelial necroptosis, tumour cell extravasation and metastasis. In contrast, pharmacological caspase inhibition or endothelial-cell-specific loss of caspase-8 promoted these processes. We furthermore show *in vitro* and *in vivo* that tumour-cell-induced endothelial necroptosis leading to extravasation and metastasis requires amyloid precursor protein expressed by tumour cells and its receptor, death receptor 6 (DR6), on endothelial cells as the primary mediators of these effects. Our data identify a new mechanism underlying tumour cell extravasation and metastasis, and suggest endothelial DR6-mediated necroptotic signalling pathways as targets for anti-metastatic therapies.

Apoptosis and necroptosis are major forms of regulated cell death[10]. Necroptosis involves in part molecules also found to regulate apoptosis but depends on the formation of the necrosome, consisting of RIPK1 and RIPK3, which activates the pseudokinase mixed lineage kinase domain-like (MLKL) by phosphorylation, leading to the execution of the necroptotic program[11–13]. Critical for the formation of the necrosome is a compromised caspase-8 activity, which functions as a mediator of apoptosis[14,15].

While studying the interaction of tumour cells (TCs) and endothelial cells (ECs), we found an increased number of dead ECs when co-cultured with TCs (Fig. 1a). Dying ECs lacked typical apoptotic features such as nuclear condensation and/or fragmentation[16] as well as annexin-V-positivity that were found in apoptotic ECs after treatment with TRAIL, staurosporine or tumour-necrosis factor-α (TNF-α) (Fig. 1a and Extended Data Fig. 1a–c). Instead, similar to nuclei of cells that underwent H_2O_2- or hypoxia-induced accidental necrosis, nuclei of ECs exposed to TCs showed minor changes in morphology but were positive for ethidium-homodimer-III (EthD-III), which, like propidium iodide, indicates necrotic cells with compromised membrane integrity[16] (Fig. 1a, Extended Data Fig. 1a, d and Supplementary Videos 1–4). TC-induced necrotic EC death was not affected by addition of peripheral blood mononuclear cells and platelets (Extended Data Fig. 1e) and was seen during co-culture of different primary human and murine ECs with a variety of TC lines in a concentration-dependent manner (Fig. 1b and Extended Data Fig. 1f–h).

EthD-III-positive ECs also co-stained for phospho-MLKL (Extended Data Fig. 1i), and knockdown of RIPK3 or MLKL in ECs as well as

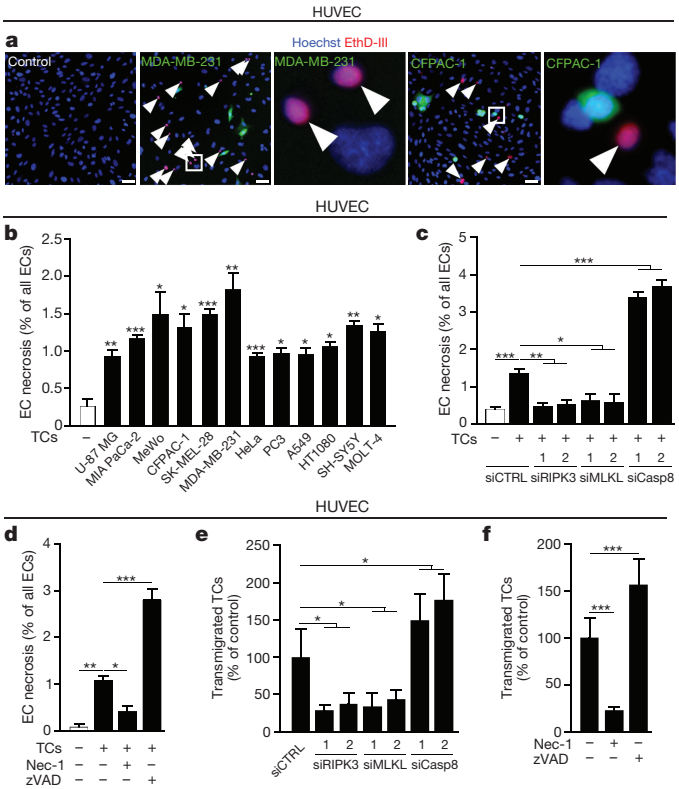


Figure 1 | TC-induced endothelial necroptosis promotes TC transendothelial migration. a, Fluorescent images of human umbilical vein ECs (HUVECs) cultured in the presence of TCs (green) and stained as indicated. Arrowheads, EthD-III-positive ECs; scale bar, 10 μm. **b**, EC necrosis upon exposure to different TCs. **c–f**, EC necrosis in the presence of MDA-MB-231 TCs (**c**, **d**) and MDA-MB-231 TC transmigration over an endothelial layer (**e**, **f**) after siRNA-mediated knockdown in ECs as indicated (**c**, **e**) or after treatment with Nec-1 (30 μM) or z-VAD-fmk (zVAD, 100 μM) (**d**, **f**). Shown are representative data of two (**b**) or three (**c–f**) independent experiments with mean values ± s.e.m. (**b–d**) or ± s.d. (**e**, **f**) from biological triplicates ($n = 3$) (**b**) or sextuplicates ($n = 6$) (**c–f**). $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$. One-way analysis of variance (ANOVA) and Bonferroni's post hoc test.

[1]Max Planck Institute for Heart and Lung Research, Department of Pharmacology, Ludwigstrasse 43, 61231 Bad Nauheim, Germany. [2]University of Cologne, Institute for Genetics, Center for Molecular Medicine (CMMC), and Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), Joseph-Stelzmann-Strasse 26, 50931 Cologne, Germany. [3]University of Heidelberg, Department of Bioinformatics and Functional Genomics, Institute for Pharmacy and Molecular Biotechnology, Im Neuenheimer Feld 364, 69120 Heidelberg, Germany. [4]J. W. Goethe University Frankfurt, Medical Faculty, Theodor-Stern-Kai 7, 60590 Frankfurt, Germany.

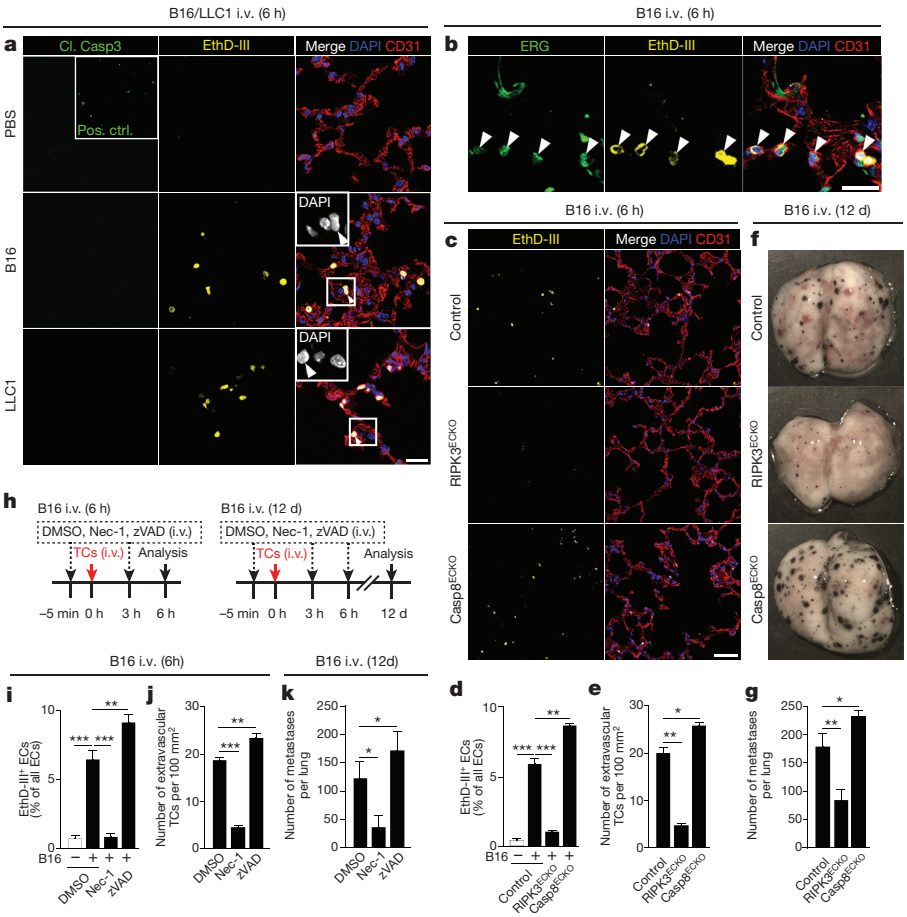


Figure 2 | TC-induced EC necroptosis and metastasis require endothelial RIPK3. **a**, **b**, Confocal images of lung sections 6 h after i.v. injection of B16 or LLC1 TCs into C57BL/6 WT animals stained for the indicated markers. Arrowheads, EthD-III-positive cells; scale bar, 20 μm. **c**–**k**, B16 TCs were i.v. injected into EC-specific RIPK3 or caspase-8 knockout mice ($RIPK3^{ECKO}$, $Casp8^{ECKO}$) (**c**–**g**) or into WT animals treated with dimethylsulfoxide (DMSO), Nec-1 or z-VAD-fmk (zVAD) (**i**–**k**). Six hours later, pulmonary EthD-III-positive ECs (**d**, **i**) and extravascular TCs (**e**, **j**) were quantified, and 12 d later lung metastases were determined (**g**, **k**). **c**, **f**, Confocal images of lung sections (scale bar, 50 μm) and images of lungs. **h**, Experimental design as performed in **i**–**k**. Cre-negative littermates served as controls (**c**–**g**). Shown are representative data of three independent experiments with mean values ± s.e.m. (**d**, **e**, **i**, **j**) or ± s.d. (**g**, **k**) from $n = 4$ (**d**, **e**, **i**, **j**) or $n = 6$ animals (**g**, **k**) per condition. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$. One-way ANOVA and Bonferroni's post hoc test.

treatment of cells with the RIPK1 kinase inhibitor necrostatin-1 (Nec-1) blocked TC-induced endothelial necrotic cell death (Fig. 1c, d and Extended Data Fig. 1j, k). In contrast, knockdown of caspase-8, a major negative regulator of necroptosis[14,15], or inhibition of caspases by treatment of cells with z-VAD-fmk (zVAD) increased the number of necrotic ECs (Fig. 1c, d and Extended Data Fig. 1j). These data identify the mode of TC-induced EC death as necroptosis. Interestingly, inhibition of TC-induced necroptosis reduced TC migration over an EC layer whereas enhanced necroptosis promoted transmigration (Fig. 1e, f).

Within a few hours after intravenous (i.v.) injection of syngeneic B16F10 (B16) melanoma or LLC1 (Lewis lung carcinoma line 1) cells into C57BL/6 wild-type (WT) mice, the lungs of these animals showed EthD-III-positive cells (Fig. 2a and Extended Data Fig. 2a), which co-stained for the EC markers ERG and CD31 (Fig. 2b and Extended Data Fig. 2b) but did not show any overlap with fluorescently labelled TCs or CD45-positive leukocytes (Extended Data Fig. 2c). EthD-III-positive ECs showed no signs of chromatin condensation or fragmentation (Fig. 2a) and were negative for cleaved caspase-3, TdT-mediated dUTP nick end labelling (TUNEL) or annexin-V (Fig. 2a and Extended Data Fig. 2d). TC-induced EC necrosis was not a result of physical blood vessel occlusion, as it was not reproduced by i.v. injection of equal amounts of 15 μm microspheres (Extended Data Fig. 2e). In mice with tamoxifen-induced endothelium-specific RIPK3-deficiency ($Tie2\text{-}CreER^{T2}$;$RIPK3^{loxP/loxP}$ ($RIPK3^{ECKO}$); Extended Data Fig. 3a, b), which had normal pulmonary vascular permeability (Extended Data Fig. 3c), numbers of EthD-III-positive ECs induced after TC injection were reduced (Fig. 2c, d). Moreover, the number of transmigrated TCs over ECs derived from these knockout animals was reduced *in vitro* (Extended Data Fig. 3d), and loss of endothelial RIPK3 expression strongly reduced numbers of extravasated TCs 6 h after i.v. TC injection as well as numbers of metastases 12 d later (Fig. 2e–g and Extended Data Fig. 3e) or of metastases derived from primary tumours (Extended Data Fig. 3f). Similar effects were observed in MLKL-deficient mice (Extended Data Fig. 4) or after short-term treatment of WT mice with Nec-1, which did not affect TC proliferation, viability or migration *in vitro* (Fig. 2h–k and Extended Data Fig. 5a–e). Also Nec-1s reduced metastasis of murine TCs (Extended Data Fig. 5f), and Nec-1 treatment resulted in fewer necroptotic ECs and metastases when human MDA-MB-231 TCs were injected into immunodeficient SCID mice (Extended Data Fig. 5g, h). In contrast, animals with induced EC-specific caspase-8-deficiency ($Tie2\text{-}CreER^{T2}$;$Casp8^{loxP/loxP}$ ($Casp8^{ECKO}$); Extended Data Fig. 6a), which had normal pulmonary vascular permeability (Extended Data Fig. 6b), and WT mice treated with zVAD showed increased numbers of EthD-III-positive ECs and extravasated TCs, and animals developed more metastases including those from primary tumours (Fig. 2c–k and Extended Data Fig. 6c–i). Thus, TCs induce necroptotic EC death also *in vivo* and EC necroptosis facilitates extravasation of TCs and promotes metastasis formation.

To identify potential endothelial receptors that mediate TC-induced necroptosis, we performed short interfering RNA (siRNA)-mediated knockdowns of 32 candidate genes (Supplementary Table 1). Knockdown of DR6 (also known as tumour-necrosis factor receptor superfamily 21 (TNFRSF21)) in ECs reduced TC-induced necroptosis as well as transendothelial TC migration (Fig. 3a–c and Extended Data Fig. 7a). The TNF receptor family member DR6 is expressed by mouse and human ECs of different vascular beds (Extended Data Fig. 7b–d), and contains a cytosolic death domain that enables cell death signalling[17,18]. Mice lacking DR6 showed reduced metastasis (Fig. 3d), and DR6 expressed by immune cells[17,19] is not involved in metastasis (Extended Data Fig. 7e–j). Blocking DR6-function by an anti-DR6 antibody (5D10)[20] protected ECs from TC-induced necroptosis and strongly reduced the number of TCs migrating over an endothelial layer (Extended Data Fig. 8a, b). The TNF-α blocker etanercept had no effect (Extended Data Fig. 8c–e). Animals treated within 6 h after TC injection with the anti-DR6 antibody showed reduced

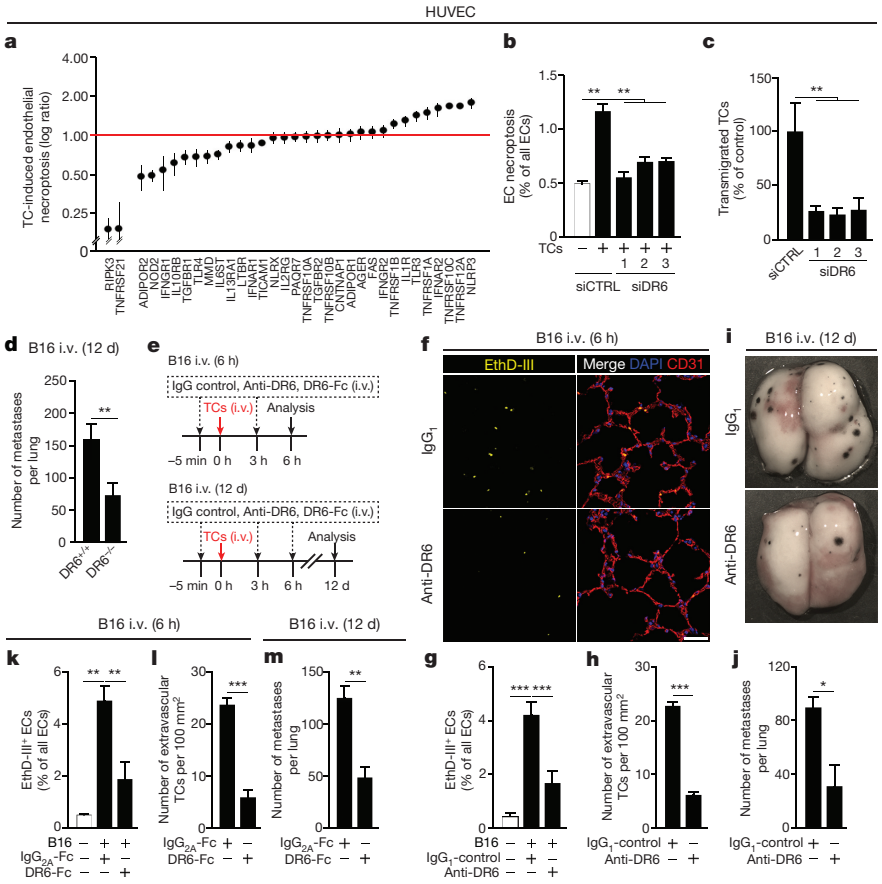


Figure 3 | DR6 mediates TC-induced endothelial necroptosis and metastasis formation. a, Effect of endothelial siRNA-mediated knockdowns on MDA-MB-231 TC-induced necroptosis. **b**, **c**, Effects of endothelial DR6 (TNFRSF21) knockdown on MDA-MB-231 TC-induced EC necroptosis (**b**) and transmigrated TCs (**c**). **d**, Lung metastases in $DR6^{-/-}$ mice 12 d after i.v. injection of B16 TCs. **e**, Experimental design as performed in **f**–**m**. **f**–**m**, C57BL/6 WT animals were treated with an anti-DR6 antibody (**f**–**j**) or the extracellular DR6-domain fused to Fc (DR6-Fc) (**k**–**m**). Thereafter, B16 TCs were injected i.v. and lungs were analysed after 6 h for pulmonary EthD-III-positive ECs (**g**, **k**) and extravascular TCs (**h**, **l**) or after 12 d for metastases (**j**, **m**). **f**, **i**, Confocal images of lung sections (scale bar, 50 μm) and images of lungs. IgG_1 antibody or Fc domain of IgG_{2A} (IgG_{2A}-Fc) served as controls. Shown are representative data of three (or two for **d**) independent experiments with mean values ± s.e.m. (**b**, **g**, **h**, **k**, **l**) or ± s.d. (**c**, **d**, **j**, **m**) from biological sextuplicates ($n = 6$) (**b**, **c**) or from $n = 4$ (**g**, **h**, **k**, **l**) or $n = 5$ animals (**d**, **j**, **m**) per condition. $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$. One-way ANOVA and Bonferroni's post hoc test (**b**, **c**, **g**, **k**) or unpaired, two-tailed Student's *t*-test (**d**, **h**, **j**, **l**, **m**).

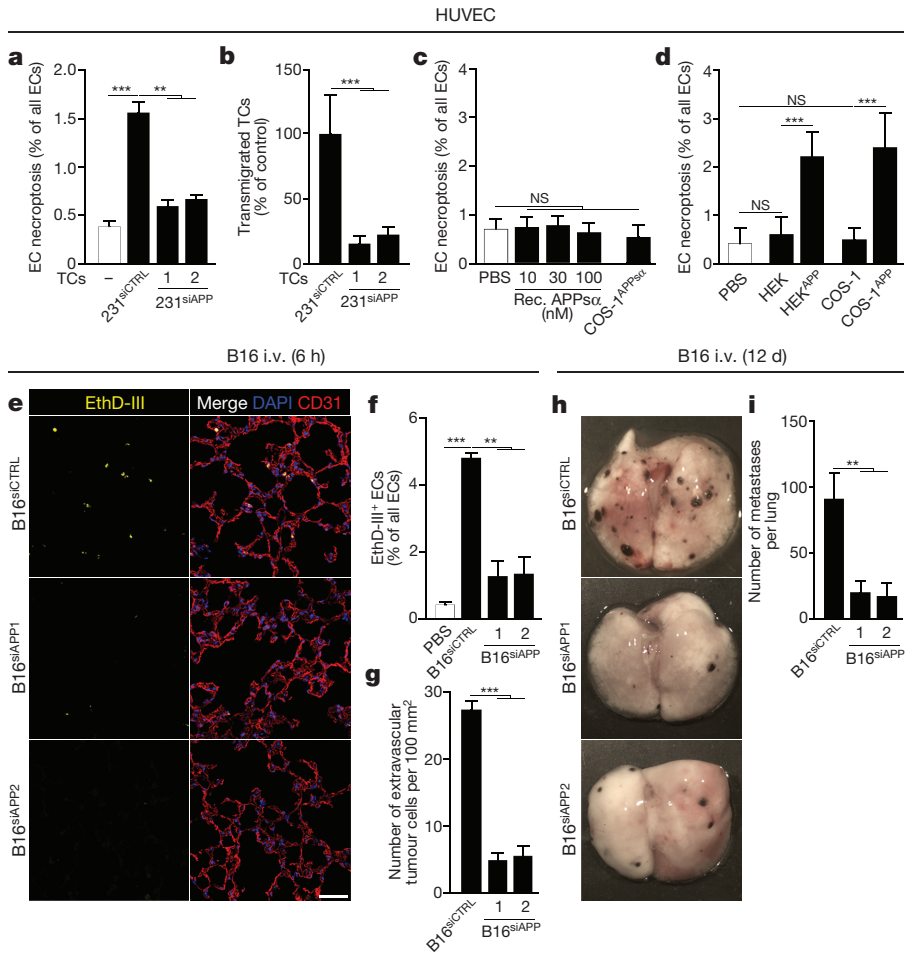


Figure 4 | APP expressed by TCs induces endothelial necroptosis and promotes metastasis. a–**d**, EC necroptosis (**a**, **c**, **d**) and TC transendothelial migration (**b**) upon co-culture with MDA-MB-231 TCs with silenced APP expression (231^{siAPP}) (**a**, **b**) or in the presence of recombinant soluble extracellular APPsα or COS-1 cells overexpressing the soluble APPsα fragment (**c**) or in the presence of HEK or COS-1 cells overexpressing membrane-bound full-length APP (HEK^{APP}, $COS\text{-}1^{APP}$) (**d**). **e**–**i**, C57BL/6 WT mice were injected i.v. with B16 TCs with unsilenced ($B16^{siCTRL}$) or with silenced APP expression ($B16^{siAPP}$), and lungs were analysed after 6 h for pulmonary EthD-III-positive ECs (**f**) and extravascular TCs (**g**) or after 12 d for lung metastases (**i**). **e**, **h**, Confocal images of lung sections (scale bar, 50 μm) and images of lungs. Shown are representative data of three independent experiments with mean values ± s.e.m. (**a**, **c**, **d**, **f**, **g**) or ± s.d. (**b**, **i**) from biological sextuplicates ($n = 6$) (**a**–**d**) or from $n = 4$ animals per condition (**f**, **g**) or $n = 8$ ($B16^{siCTRL}$), $n = 7$ ($B16^{siAPP\#1}$) or $n = 5$ ($B16^{siAPP\#2}$) animals (**i**). $^{**}P < 0.01$; $^{***}P < 0.001$; NS, not significant. One-way ANOVA and Bonferroni's post hoc test.

numbers of necroptotic ECs and extravasated TCs, and these animals developed fewer metastases (Fig. 3e–j and Extended Data Fig. 8f). These DR6-mediated effects required ligand binding to DR6 as the DR6 ectodomain fused to an Fc fragment (DR6-Fc) inhibited TC-induced endothelial necroptosis and transendothelial migration of TCs as well as development of metastasis (Fig. 3e, k–m and Extended Data Fig. 8g–j).

A previously identified ligand of DR6 is amyloid precursor protein (APP)[21,22] which is widely expressed including in various TCs[23–25] (Extended Data Fig. 9a, b). Knockdown of APP expression in TCs strongly reduced their ability to induce endothelial necroptosis and to transmigrate an EC layer (Fig. 4a, b and Extended Data Fig. 9c). Neither conditioned media from TC–EC co-cultures (Extended Data Fig. 9d) nor the purified soluble APPsα fragment[26] or cells transfected with constructs to overexpress APPsα-induced endothelial necroptosis (Fig. 4c and Extended Data Fig. 9e). However, necroptosis was induced by direct contact of ECs with cells overexpressing the transmembrane full-length form of APP (HEK^{APP} or $COS\text{-}1^{APP}$) (Fig. 4d and Extended Data Fig. 9e). Thus, consistent with the notion that the efficacy of TNF-receptor family ligands to induce cell death is strongly increased when present in their transmembrane form[27], APP exposed by TCs rather than soluble APP released from TCs is required for TC-induced DR6-mediated endothelial necroptosis. TCs in which APP expression was strongly reduced by siRNA-mediated knockdown showed normal proliferation, cell survival and basal migratory activity (Extended Data Fig. 10a–d) but reduced ability to transmigrate a layer of primary lung ECs (Extended Data Fig. 10e, f). In addition, these TCs almost completely lost the ability after i.v. injection to induce EC necroptosis, to extravasate and to form metastases (Fig. 4e–i and Extended Data Fig. 10g). Interestingly, epidemiological data indicate that high APP expression in TCs is associated with higher frequency of metastasis formation and poor prognosis[23–25].

How does endothelial necroptosis facilitate metastasis formation? It is conceivable that dying ECs provide a gap through which TCs can pass and start to extravasate (Extended Data Fig. 10h). It is also possible that damage-associated molecular pattern (DAMP) molecules, which are released from lysed necroptotic cells[28], act on TCs, neighbouring ECs or other cells and thereby promote TC extravasation and metastasis (Extended Data Fig. 10h). A DAMP molecule associated with necrotic or necroptotic cell death is ATP, which was shown to have pro-migratory effects on TCs[29] and to induce opening of the endothelial barrier and to promote TC metastasis[30].

Our data show that TCs induce endothelial necroptotic death via APP and DR6, which is required for efficient TC extravasation and metastases. Since necrosis or necroptosis are rare events under physiological conditions, targeting DR6-mediated necroptosis of ECs in the context of tumour progression may represent a novel approach to prevent or treat metastasis.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 25 August 2015; accepted 4 July 2016.
Published online 3 August 2016.

1. Valastyan, S. & Weinberg, R. A. Tumor metastasis: molecular insights and evolving paradigms. *Cell* **147,** 275–292 (2011).
2. Wan, L., Pantel, K. & Kang, Y. Tumor metastasis: moving new biological insights into the clinic. *Nature Med.* **19,** 1450–1464 (2013).
3. Vanharanta, S. & Massagué, J. Origins of metastatic traits. *Cancer Cell* **24,** 410–421 (2013).
4. Bos, P. D. *et al.* Genes that mediate breast cancer metastasis to the brain. *Nature* **459,** 1005–1009 (2009).
5. Padua, D. *et al.* TGFβ primes breast tumors for lung metastasis seeding through angiopoietin-like 4. *Cell* **133,** 66–77 (2008).
6. Gupta, G. P. *et al.* Mediators of vascular remodelling co-opted for sequential steps in lung metastasis. *Nature* **446,** 765–770 (2007).
7. Reymond, N., d'Água, B. B. & Ridley, A. J. Crossing the endothelial barrier during metastasis. *Nature Rev. Cancer* **13,** 858–870 (2013).
8. Labelle, M. & Hynes, R. O. The initial hours of metastasis: the importance of cooperative host-tumor cell interactions during hematogenous dissemination. *Cancer Discov.* **2,** 1091–1099 (2012).
9. Joyce, J. A. & Pollard, J. W. Microenvironmental regulation of metastasis. *Nature Rev. Cancer* **9,** 239–252 (2009).
10. Galluzzi, L. *et al.* Essential versus accessory aspects of cell death: recommendations of the NCCD 2015. *Cell Death Differ.* **22,** 58–73 (2015).
11. Pasparakis, M. & Vandenabeele, P. Necroptosis and its role in inflammation. *Nature* **517,** 311–320 (2015).
12. Zhou, W. & Yuan, J. Necroptosis in health and diseases. *Semin. Cell Dev. Biol.* **35,** 14–23 (2014).
13. Silke, J., Rickard, J. A. & Gerlic, M. The diverse role of RIP kinases in necroptosis and inflammation. *Nature Immunol.* **16,** 689–697 (2015).
14. Oberst, A. *et al.* Catalytic activity of the caspase-8–$FLIP_L$ complex inhibits RIPK3-dependent necrosis. *Nature* **471,** 363–367 (2011).
15. Kaiser, W. J. *et al.* RIP3 mediates the embryonic lethality of caspase-8-deficient mice. *Nature* **471,** 368–372 (2011).
16. Krysko, D. V., Vanden Berghe, T., D'Herde, K. & Vandenabeele, P. Apoptosis and necrosis: detection, discrimination and phagocytosis. *Methods* **44,** 205–221 (2008).
17. Pan, G. *et al.* Identification and functional characterization of DR6, a novel death domain-containing TNF receptor. *FEBS Lett.* **431,** 351–356 (1998).
18. Lavrik, I., Golks, A. & Krammer, P. H. Death receptor signaling. *J. Cell Sci.* **118,** 265–267 (2005).
19. Liu, J. *et al.* Enhanced $CD4^+$ T cell proliferation and Th2 cytokine production in DR6-deficient mice. *Immunity* **15,** 23–34 (2001).
20. Huang, G. *et al.* Death receptor 6 (DR6) antagonist antibody is neuroprotective in the mouse $SOD1^{G93A}$ model of amyotrophic lateral sclerosis. *Cell Death Disease* **4,** e841 (2013).
21. Nikolaev, A., McLaughlin, T., O'Leary, D. D. M. & Tessier-Lavigne, M. APP binds DR6 to trigger axon pruning and neuron death via distinct caspases. *Nature* **457,** 981–989 (2009).
22. Xu, K., Olsen, O., Tzvetkova-Robev, D., Tessier-Lavigne, M. & Nikolov, D. B. The crystal structure of DR6 in complex with the amyloid precursor protein provides insight into death receptor activation. *Genes Dev.* **29,** 785–790 (2015).
23. Takagi, K. *et al.* Amyloid precursor protein in human breast cancer: an androgen-induced gene associated with cell proliferation. *Cancer Sci.* **104,** 1532–1538 (2013).
24. Takayama, K. *et al.* Amyloid precursor protein is a primary androgen target gene that promotes prostate cancer growth. *Cancer Res.* **69,** 137–142 (2009).
25. Yang, Z., Fan, Y., Deng, Z., Wu, B. & Zheng, Q. Amyloid precursor protein as a potential marker of malignancy and prognosis in papillary thyroid carcinoma. *Oncol. Lett.* **3,** 1227–1230 (2012).
26. Hick, M. *et al.* Acute function of secreted amyloid precursor protein fragment APPsα in synaptic plasticity. *Acta Neuropathol.* **129,** 21–37 (2015).
27. O'Reilly, L. A. *et al.* Membrane-bound Fas ligand only is essential for Fas-induced apoptosis. *Nature* **461,** 659–663 (2009).
28. Kaczmarek, A., Vandenabeele, P. & Krysko, D. V. Necroptosis: the release of damage-associated molecular patterns and its physiological relevance. *Immunity* **38,** 209–223 (2013).
29. Takai, E. *et al.* Autocrine regulation of TGF-β1-induced cell migration by exocytosis of ATP and activation of P2 receptors in human lung cancer cells. *J. Cell Sci.* **125,** 5051–5060 (2012).
30. Schumacher, D., Strilic, B., Sivaraj, K. K., Wettschureck, N. & Offermanns, S. Platelet-derived nucleotides promote tumor-cell transendothelial migration and metastasis via $P2Y_2$ receptor. *Cancer Cell* **24,** 130–137 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank S. Fulda and our friends and colleagues for comments on the manuscript. We also thank S. Hümmer for secretarial help and C. Ringel, J. Hoffmann, I.-M. Gross, D. Magalei and M. Winkels for technical help. This work was supported by the German Cancer Aid and the Max Planck Society. K.H. was supported by the China Scholarship Council. U.C.M. was supported by a grant from the Deutsche Forschungsgemeinschaft (MU 1457/9-2). M.P. received funding from the European Research Council (grant agreement 323040), the Deutsche Forschungsgemeinschaft (SFB670, SFB829), Worldwide Cancer Research (grant 15-0228) and the Helmholtz Alliance Preclinical Comprehensive Cancer Center.

Author Contributions B.S. performed most of the experiments and analysed the data. L.Y. generated mice with a conditional *Ripk3* allele and contributed to *in vitro* and *in vivo* studies. J.A.J. contributed to *in vitro* experiments. L.W. and M.P. generated MLKL-deficient animals. K.H. and U.C.M. purified APPsα and performed APP-related experiments. B.S. and S.O. designed the study, discussed data and wrote the manuscript. All authors commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.S. (boris.strilic@mpi-bn.mpg.de) or S.O. (stefan.offermanns@mpi-bn.mpg.de).

Reviewer Information *Nature* thanks C. Betsholtz, S. Tavazoie and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Cell death assays. HUVECs, human microvascular vein ECs from lung (HMVECs-L) or L929 cells (1.5×10^4 at seeding in 100 µl) were cultured for 24 h in 96-well plates. To induce cell death, cells were stimulated overnight with rhTRAIL (100 ng/ml, Peprotech), rhTNF-α (50 ng/ml, Peprotech), Staurosporine (0.5 µM, Jena BioScience), H_2O_2 (1 mM, AppliChem) or rmTNF-α (100 ng/ml, Peprotech) or cultured under hypoxic conditions (1% O_2, O_2 Control Glove Boxes, Coy Laboratory). Alternatively, for co-culture experiments, 1.5×10^3 green fluorescent protein (GFP)-expressing TCs, calcein-AM-labelled TCs, COS-1 or HEK293T cells (for more details see below), or freshly isolated human peripheral blood mononuclear cells (PBMCs) stained with calcein-AM containing 20 times the number of platelets were added alone, in combination with each other or in the presence of the indicated substances onto the EC monolayer and cultured overnight: Nec-1 (30 µM), z-VAD-fmk (100 µM), etanercept (20 µg/ml), DR6-Fc or IgG_1-Fc (0.1–1 µg/ml), antagonistic antibody against DR6 (5D10) or IgG_1 isotype control antibody (1–30 µg/ml). Unless stated otherwise, HUVECs and MDA-MB-231-GFP TCs were used for *in vitro* studies. PBMCs were isolated using standard protocols with Ficoll density gradient centrifugation. For supernatant experiments, HUVEC monolayers grown to confluency were cultured with conditioned medium obtained from HUVECs co-cultured in the presence of TCs for 18 h. For knockdown experiments, 1.5×10^4 HUVECs were transfected using Lipofectamine RNAiMAX (Life Technologies) with different sets of siRNAs (Sigma or Qiagen, see Supplementary Table 1) and cultured on 96-well plates. In cases where siRNA-mediated knockdown was performed on TCs, cells were transfected using Lipofectamine RNAiMAX with different sets of siRNA (Sigma) and seeded 48 h after transfection on confluent monolayers of ECs. Knockdown efficiencies were determined by western blotting upon lysis with Laemmli buffer or by quantitative RT–PCR (LightCycler480, Roche). TC number upon gene knockdown or after treatment with Nec-1 or zVAD was determined by counting Hoechst 33342-positive cells, and TC death was determined by counting condensed and/or EthD-III-positive nuclei (see below). Cell migration was determined by a scratch assay[31].

Cell death analysis. For all conditions, EthD-III (1.6 µM, Biotium) and Hoechst 33342 (2 µM, Thermo Scientific) were added shortly before automated image acquisition in an atmosphere-controlled chamber (37 °C, 5% CO_2) using an Olympus IX81 microscope or before overnight culture. On the basis of cells cultured under defined apoptotic, necrotic or necroptotic conditions and stained with Hoechst 33342 (a cell-permeable nuclear dye) and EthD-III (a membrane-impermeant nuclear dye), morphological criteria for discriminating apoptotic from necrotic (or necroptotic) cells compared with living cells were defined as follows. A living cell has a normal round to kidney-shaped nucleus (as visualized by Hoechst 33342) and is negative for EthD-III. An apoptotic cell has a strongly condensed and fragmented nucleus and is negative for EthD-III. A necrotic or necroptotic cell has a normal round to kidney-shaped nucleus or a minor degree of nuclear shrinkage (no condensation and no fragmentation) and is positive for EthD-III. A late apoptotic cell is positive for EthD-III but can be discriminated from a necrotic/necroptotic cell on the basis of the strong condensation and fragmentation of its nucleus. To prevent repeatedly counting fragmented parts of apoptotic nuclei, a Gaussian blur with a radius of three pixels was applied to all images to be analysed. ECs were defined as GFP- or calcein-AM-negative cells. The total number of all endothelial nuclei was determined through a low threshold (TH1) and application of a watershed on the resulting binary image over all Hoechst 33342-positive nuclei (minus the nuclei from TCs). When possible, a second separate threshold (TH2) was used to determine the number of condensed nuclei. The number of EthD-III-positive cells was determined through an independent second low threshold only. In cases where this automated analysis failed, the mode of cell death was determined manually for each individual EC by application of the criteria summarized above. All images were analysed in ImageJ (National Institutes of Health). Unless stated otherwise, each experiment was performed at least three times with a minimum of six wells per condition and four independent images acquired per well.

Transwell assays. Assays were performed as described previously[30]. Briefly, ECs (1.5×10^4 at seeding in 50 µl) were cultured for 2 d or, for knockdown experiments, 8×10^3 HUVECs were transfected using Lipofectamine RNAiMAX with different sets of siRNA (Sigma or Qiagen, see Supplementary Table 1) and cultured on 96-transwell plates with polyester membranes of 8 µm pore size (Corning) with daily medium changes until reaching confluency. For transmigration experiments, the medium of the upper compartment was removed and 7.5×10^3 GFP-expressing or calcein-AM-labelled TCs were added in 50 µl endothelial culture medium alone or in the presence of different substances (see above). For all experiments, transmigrated TCs on the lower side of the filter were imaged (Zeiss Axio Observer. Z1 or Olympus IX81) and quantified with ImageJ. Each experiment was performed at least three times with a minimum of six wells per condition.

Metastasis models. Unlabelled or CFSE-labelled TCs (B16F10 melanoma or LLC1 lung carcinoma cells) or fluorescent microspheres (15 µm diameter, Life Technologies) in 50 µl PBS were injected to the lateral tail vein of mice at 1,600 TCs or microspheres per gram body weight. For inhibitory experiments, two doses of 25 µl Nec-1 (1.65 µg/g), Nec-1s (1.65 µg/g), z-VAD(OMe)-fmk (4 µg/g), etanercept (10 mg/kg), rmDR6-Fc (0.2 µg/g), $rmIgG_{2A}$-Fc (0.2 µg/g), anti-DR6 (3.5 µg/g) or control IgG_1 (3.5 µg/g) were injected into the tail vein shortly before and 3 h after TC injection. For evaluation of TC-induced EC death, 6 h after injection of TCs, 50 µl EthD-III (300 µM in PBS) were injected i.v. and after 10 min animals were killed and perfused with PBS and 4% paraformaldehyde and directly processed for immunohistochemical analysis. For evaluation of the number of extravasated TCs, CFSE-labelled B16 TCs were injected i.v. and 6 h later non-perfused lungs were isolated and fixed in 4% paraformaldehyde. Stained cryosections were analysed in *xyz* views on a Leica SP5 confocal microscope. The number of EthD-III-positive ECs was determined by manual counting of EthD-III/ERG- or EthD-III/CD31-double-positive cells on central areas of a minimum of four random longitudinal sections of the left lobe and the superior and middle right lobe of each lung. For quantification of extravasated TCs, cryosections were analysed by two criteria: TCs directly surrounded by CD31 staining (that is, blood vessel) and with a non-invasive phenotype (that is, round cell shape) were scored as intravascular, while cells outside blood vessels with an invasive phenotype (that is, irregular cell shape with protrusions) were scored as extravascular. For both analyses, numbers of EthD-III-positive ECs or extravasated TCs from each section of one lung were averaged (average per mouse), and these values were again averaged to obtain a mean value ± s.e.m. from one experiment. For evaluation of lung metastases, an additional (third) treatment with the aforementioned substances was performed at 6 h after TC i.v. injection, and lung metastases were analysed 12 d later either macroscopically (B16/LLC1 i.v.) or microscopically (MDA-MB-231 i.v.) by analysing every tenth section after whole-lung sectioning and H&E staining. Alternatively, to study metastasis formation from primary tumours, 1×10^6 B16 or LLC1 TCs in 25 µl PBS were injected subcutaneously into the shaved flank of mice, and the primary tumour was surgically removed 10 d later, followed by microscopic lung metastases analysis 27 d later as described above. The mean number of lung metastases per experiment was determined by averaging the individual values of lung metastases from each lung resulting in a mean value ± s.d. from one experiment. A minimum of three animals per group was used. Animal groups were sex-matched, and mice were 8–16 weeks of age. Mice were grouped randomly. Experiments were performed blind for genotype of the mice. All experimental animal procedures were approved by the Hessian Regional Board.

Mice. Control C57BL/6 and SCID animals were obtained from Charles River; B6F2D1 animals were obtained from Harlan. $DR6^{-/-}$ animals were provided by Genentech (USA). Caspase-$8^{loxP/loxP}$ animals were a gift from S. Hedrick (University of California San Diego, USA) provided by C. Becker (University Medical Center Erlangen, Germany) and crossed to tamoxifen-inducible Tie2-$CreER^{T2}$ animals[32] to obtain Tie2-$CreER^{T2}$;Caspase-$8^{loxP/loxP}$ animals ($Casp8^{ECKO}$). To generate RIPK3 conditional knockout animals, an 880-bp fragment containing loxP-flanked exon 2 and 3 from *Ripk3* (to obtain a premature stop codon upon cre-mediated recombination) as well as the 5′ homology arm and the 3′ homology arm was amplified from BAC RPCI-23-237G18 (Children's Hospital Oakland Research Institute) and cloned into the pKOII targeting vector additionally containing a Frt-flanked neomycin resistance gene (neo^R) and the diphtheria toxin A gene (*dta*) as negative selection marker. The targeting vector was linearized with NotI and introduced into V6.5 ES cells by electroporation. Upon treatment with 400 µg/ml G418, DNA from 400 clones was isolated, and screened for correct recombination by Southern blot. Two independent ES cell clones were injected into C57BL/6 blastocysts, which were subsequently transferred to pseudo-pregnant females to generate chimaeric offspring. Male chimaeras were bred with C57BL/6 female mice to produce heterozygotes. The germ line transmission was confirmed in the F_1 generation using a PCR genotyping strategy. Mice were then crossed to Flp-deleter mice to remove the neomycin cassette and thereafter crossed with Tie2-$CreER^{T2}$ animals to obtain Tie2-$CreER^{T2}$;$RIPK3^{loxP/loxP}$. A loxP-PCR reaction was used for detection of the WT allele (+) and the floxed (fl) allele using the primers RIPK3-fl_fwd (5′-GATCCAGAGCTCCACGCCAAG-3′) and RIPK3-fl_rev (5′-TGGAGGACCAGAGGGAAGGT-3′) resulting in band sizes of 295 bp (WT) and 340 bp (fl), respectively. To induce recombination, animals were treated with 1 mg/d tamoxifen (Sigma) for 5 consecutive days and 7–9 d later experiments were started. Caspase-8 and RIPK3 deletion in ECs was confirmed by comparing protein levels on isolated ECs from lungs of induced knockout animals (Tie2-$CreER^{T2}$;Caspase-$8^{loxP/loxP}$ (= $Casp8^{ECKO}$) or Tie2-$CreER^{T2}$;$RIPK3^{loxP/loxP}$ (= $RIPK3^{ECKO}$)) with ECs from lungs of Cre-negative control animals (Caspase-$8^{loxP/loxP}$ or $RIPK3^{loxP/loxP}$) using western blot. Quantification of *in vivo* permeability was performed using the Miles assay. Briefly, 11 d after induction of the knockout by tamoxifen, mice received a 100 µl tail vein injection of 0.5% Evans blue dye in PBS.

After 30 min mice were killed, and extravasated blue dye was eluted from the lungs with formamide at 56 °C and measured by spectrometry at 620 nm. Lungs cultured *ex vivo* for 18 h after i.v. injection of 1 ml staurosporine (10 μM) served as positive controls for stainings with antibodies against cleaved caspase-3 or annexin-V or with TUNEL assay. To generate MLKL$^{-/-}$ animals a TALEN pair targeting the second exon of the *Mlkl* gene was designed using the TAL Effector Nucleotide Targeter 2.0 (https://tale-nt.cac.cornell.edu/node/add/talen). A pair targeting ~70 bp downstream of the ATG with a 19 bp spacer was chosen (*Mlkl*_Talen_l: 5′-GCCGGAAACAATGCCAGCGT-3′ and *Mlkl*_Talen_r: 5′-GCCTGCTACAGCCTCTCCAG-3′). Corresponding repeat-variable di-residues (RVDs) (NH HD HD NH NH NI NI NI HD NI NI NG NH HD HD NI NH HD NH NG for Mlkl_Talen_l and HD NG NH NH NI NH NI NH NH HD NG NH NG NI NH HD NI NH NH HD for Mlkl_Talen_r) were cloned into the RCIscript-GoldyTALEN (Addgene 38142) using the Golden Gate TALEN and TAL Effector Kit 2.0 (Addgene 1000000024)[33] for subsequent *in vitro* transcription (mMESSAGE mMACHINE T3 Transcription Kit, Ambion). mRNA was purified using RNA spin columns (NucleoSpin RNA, Macherey Nagel) and microinjected into fertilized oocytes from B6D2F1 mice at a concentration of 25 or 50 ng/μl. Surviving embryos were transferred into pseudo-pregnant fosters the next day. Tails from the F_0 generation were analysed by PCR (primers type_fwd 5′-GTCTTGCACGGTGGAGGTAT-3′ and type_rev 5′-CCCAGACGTCTCTCAGCTTC-3′ resulting in a 440 bp product) followed by T7 endonuclease I assay (T7EI; T7 endonuclease I, NEB) to detect mutant offspring. Mutant mouse 64 was analysed by sequencing (GATC-Biotech) and the 8 bp deletion (Δ8) allele, which was predicted to result in a knockout was bred to homozygosity. Homozygous mice did not show an obvious phenotype and reproduced normally. Spleens and lungs of homozygous MLKL mutant mice were analysed for MLKL protein expression by western blot (MLKL: AP14272b, Abgent; α-tubulin: T6074, Sigma-Aldrich). For genotyping of MLKL$^{-/-}$ animals the type_rev primer (see above) in combination with MLKL_TAL_wt 5′-ATGCCAGCGTCTAGGAAACC-3′ or MLKL_TAL_mut 5′-GGAAACAATGCCAGCGTCAGT-3′ were used to obtain a 245 bp WT or 244 bp knockout band, respectively.

Identification of endothelial receptors involved in the regulation of TC-induced EC death. First, expression levels of 132 genes encoding members of receptor families known to mediate different forms of programmed cell death were analysed using cDNA from HUVEC, HMVEC-L or mouse lung ECs (MLECs) (Supplementary Table 1). RNA isolation and cDNA transcription were performed using standard protocols (Qiagen, Roche) and quantitative PCR was performed using the Universal ProbeLibrary System Technology (Roche). Genes were considered to be expressed when the corresponding cDNA was at a level $>10^{-6}$-fold compared with the cDNA level of GAPDH. Genes for further screening were selected on the basis of the following two criteria: first, the gene is highly expressed in at least one of the three tested ECs (among the top 22 genes tested); second, the gene is expressed in HUVECs (to be able to perform the screen), resulting in a total of 44 genes for testing. The screen to identify potential receptors that mediate TC-induced EC death was performed in 96-well format. HUVECs were transfected using siRNA from Sigma for the initial screen. Only siRNAs resulting in expression levels <25% as determined by quantitative PCR were used. Gene knockdowns in HUVECs cultured in the absence of TCs resulting in monolayers of less than 75% confluency compared with HUVECs transfected with scrambled control-siRNA were not included in the final analysis. Individual experiments where the knockdown of RIPK3 (used as positive control) did not result in a decreased ratio of TC-induced EC death compared with cells transfected with scrambled siRNA were not included in the final analysis. Forty-eight hours after gene silencing, 2×10^3 MDA-MB-231-GFP TCs were seeded on confluent monolayers of HUVECs in the presence of 100 μM z-VAD-fmk. After overnight co-culture, the number of dead ECs was determined as described above. The screen was performed in five independent rounds with duplicates in each round. The ratio of EC necroptosis for each condition was defined as the effect of each siRNA to alter TC-induced endothelial necroptosis compared with control HUVECs (scrambled siRNA) treated with TCs. Background cell death in HUVECs transfected with scrambled siRNA and cultured without TCs was determined for each plate individually and subtracted from each value. For values >1, gene knockdown resulted in an increase in TC-induced endothelial necroptosis; for values = 1, gene knockdown resulted in no change; for values <1, gene knockdown resulted in a decrease in TC-induced endothelial necroptosis. Independent siRNAs from Qiagen were used for testing potential hits in independent experiments.

Irradiation procedure and bone marrow transplantation. DR6$^{+/+}$ WT or DR6$^{-/-}$ mice at the age of 8–10 weeks were subjected to total body irradiation with 8.5 Gy. The X-ray machine (RadSource RS-2000) was operated at 160 kV, 4.2 kW at a dose rate of 1.2 Gy/min in air. Bone marrow cells were harvested by gently flushing both femora and both tibiae with DMEM (GIBCO). Cells were centrifuged, resuspended in HBSS (without Mg^{2+} or Ca^{2+}), and 5×10^7 cells in 100 μl (viability >95%) were injected i.v. into irradiated recipient mice on the day of irradiation. TCs were injected 6 weeks after transplantation, and numbers of EC deaths and extravasated TCs were determined 6 h later. The number of surface metastases was determined after 12 days.

Generation of APP-expressing cells. To generate HEK cells stably expressing membrane-bound full-length huAPP$_{695}$ (HEK^{APP}), HEK293T cells were transfected using a modified FUGW vector[34] expressing huAPP$_{695}$ with an amino (N)-terminal myc-tag (mycAPP) under control of the ubiquitin promoter and a puromycin resistance marker for the selection. For the generation of COS-1 cells expressing the membrane-bound full-length form of huAPP$_{695}$ (COS-1^{APP}) or the soluble huAPP$_{695}$-α fragment (COS-1^{APPsα}), COS-1 cells were transiently transfected using Lipofectamine 2000 and the plasmids pCAX APP 695 or pCAX APPs-695-α, respectively. Cells were used 24 h after transfection for further analyses. For the detection of APP, cells were lysed in CHAPS Lysis Buffer or media were conditioned for 24 h and further processed for western blotting using anti-APP (MAB348, Millipore (clone 22C11)). β-Tubulin (MAB3408, Millipore) served as loading control. pCAX APP 695 and pCAX APPs-695-α were a gift from D. Selkoe and T. Young-Pearse (Addgene plasmids 30137 and 30147)[35].

Expression analyses. Cells from individual wells of a 6- or 96-well plate were harvested (1.5×10^5 or 5×10^3 at seeding) and RNA was isolated and transcribed into cDNA according to the manufacturers' protocols (Qiagen, Roche). For quantitative RT–PCR using the Roche LightCycler480 Probes Master System, 30 ng cDNA per reaction were used. Primers were designed with the online tool provided by Roche, and only primer pairs from the top three results were chosen. Relative expression levels were obtained by normalizing with GAPDH. For single-cell gene expression analyses, RNA isolation and cDNA synthesis was performed using the 96-well C1 Single-Cell Auto Prep Array for mRNA Seq (17–25 μm) from Fluidigm. qRT–PCR was performed with the LightCycler480 Probes Master System (Roche) using intron-spanning primers. Relative expression values from individual cells were determined by the $2^{(LoDct-ct)}$ method with a limit of detection ct (LoDct) value set to 35 cycles, and values were normalized to the cell with the highest gene expression (100%). Values from empty wells or wells that contained more than one cell, as determined by visual inspection, were excluded from the final analysis.

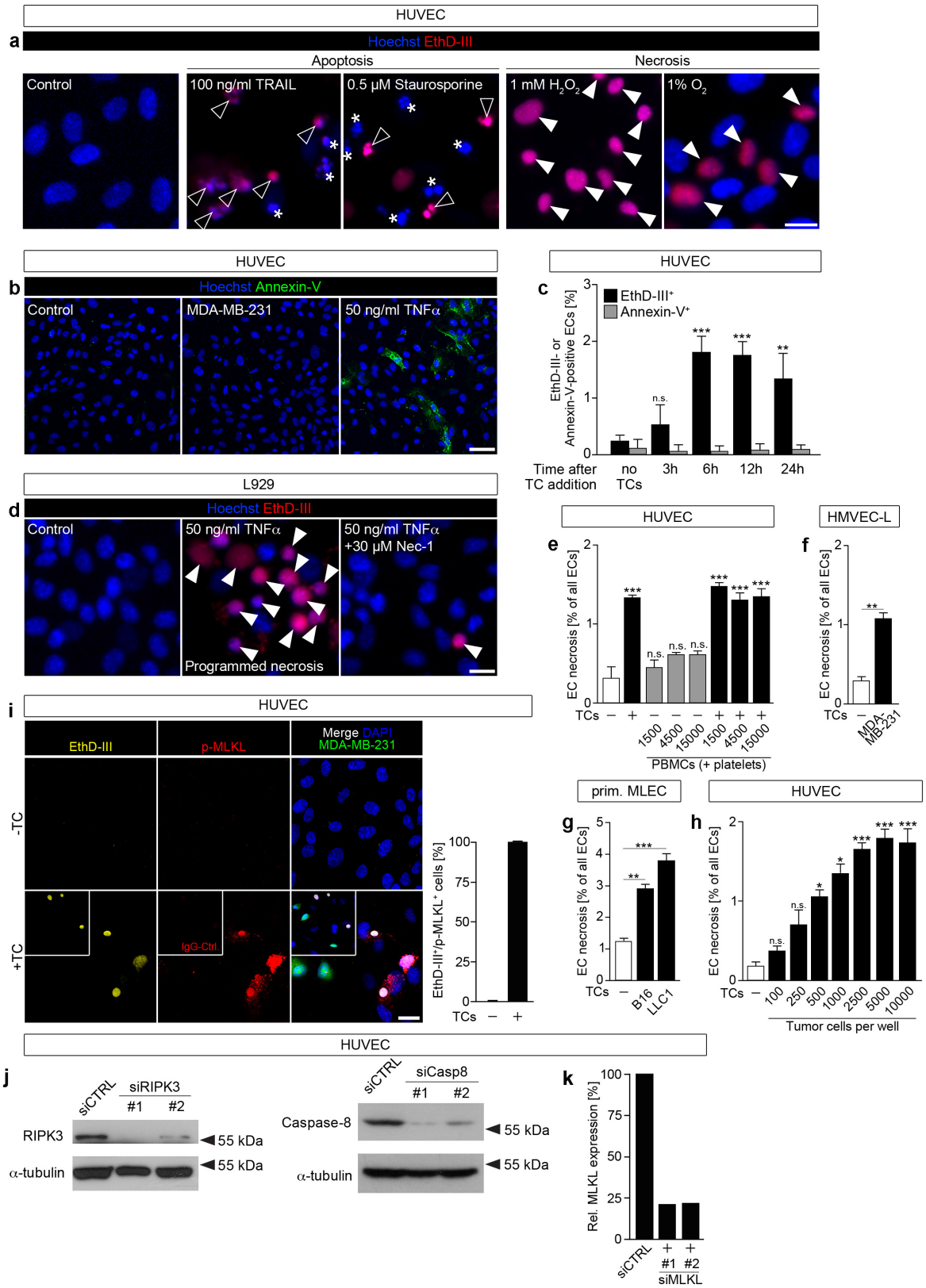
Cells. Human primary ECs and media were from Lonza. MDA-MB-231-GFP TCs were from AntiCancer. THP-1, A549, PC3, MeWo and SK-MEL-28 cells were from CLS. COS-1, B16F10 and LLC1 were from ATCC. L929 cells were a gift from J. Wiegers (Biocentre Innsbruck, Austria), U-87 MG cells were from S. Rieken (University Hospital, Heidelberg, Germany), MIA PaCa-2 and CFPAC-1 cells were from N. Giese (University Hospital, Heidelberg, Germany), Sh-SY5Y, HeLa and HT1080 were from M. Bähr (DKFZ, Germany) and MOLT-4 cells were from J. Witkowski (Medical University of Gdansk, Poland). All cells were incubated at 37 °C and 5% CO_2. HUVECs and HMVECs-L were cultured in growth factor-supplemented EGM2 or EGM2-MV medium, respectively, and passages <P6 were used for all experiments. All other cell lines were cultured in either RPMI or DMEM supplemented with 10% FBS, penicillin/streptomycin (100 units/ml) and glutamine (2 mM). Cells were tested negative for mycoplasma contamination before experiments. The cell lines are not listed in the International Cell Line Authentication Committee (ICLAC) database. Primary mouse lung ECs were isolated and cultured as described previously[36]. For knockout induction *in vitro*, cells were treated with 4-OH tamoxifen (1 μM) for 48 h.

Other materials. Media and supplements were from Life Technologies. The following reagents were used: tamoxifen (Sigma), 4-OH tamoxifen (Sigma), Nec-1 (Enzo), Nec-1s (BioVision), z-VAD-fmk (Alexis), z-VAD(OMe)-fmk (Cayman), 1-MT (Sigma), etanercept (Pfizer), rhDR6-Fc (144-DR, R&D), rhIgG$_1$-Fc (110-HG, R&D), rmDR6-Fc (6985-DR, R&D), rmIgG$_{2A}$-Fc (4460-MG, R&D), calcein-AM (AAT Bioquest), CFSE (Alexis), DAPI (Life Technologies), TUNEL staining kit (Roche). The following antibodies were used for western blot analyses: RIPK3 (ab56164, Abcam), caspase-8 (3473, ProSci), DR6 (7678R, Bioss), VE-cadherin (sc-6458, Santa Cruz), α-tubulin (T9026, Sigma). The following antibodies were used for immunohistochemistry: anti-CD31 (human) (NB600-562, Novus), anti-CD31 (mouse) (550274, BD Biosciences), anti-DR6 (7678R, Bioss), anti-cleaved caspase-3 (9661S, Cell Signaling), anti-annexin-V (sc-4252, Santa Cruz), anti-ERG (ab110639, Abcam), anti-CD45 (553082, BD Biosciences), anti-phospho-MLKL (ab187091, Abcam). Antagonistic anti-DR6 (clone 5D10) and control IgG$_1$ (clone MOPC-21) antibodies were provided by S. Mi (Biogen Idec, Cambridge, USA). Recombinant APPsα was purified as described previously[26].

Human samples. Frozen human tissue samples were obtained from Zyagen. Experiments with human samples were performed according to the regulations of the local ethics committee of the Hessian Regional Medical Board, and informed consent was obtained from all subjects.

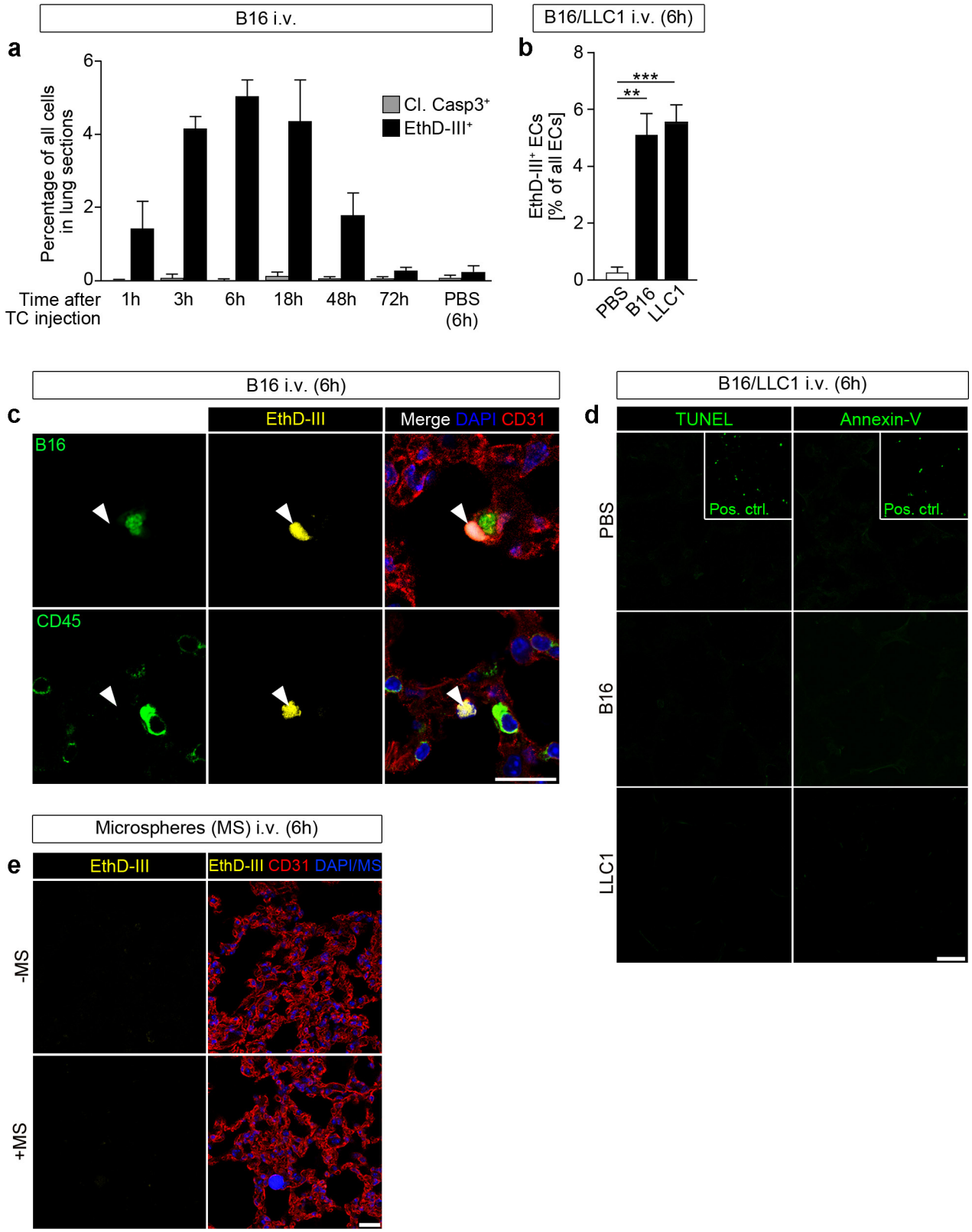
Statistical analysis. Trial experiments or experiments done previously were used to determine sample size with adequate statistical power. *In vitro* experiments were not randomized and the investigators were not blinded to them, whereas the *in vivo* experiments were randomized and investigators were blinded. Samples were excluded in cases where RNA/cDNA quality or tissue quality after processing was poor (below commonly accepted standards). Animals were excluded from experiments if they showed any signs of sickness. Data represent biological replicates. In all studies, comparison of mean values was conducted with unpaired, two-tailed Student's *t*-test or one-way or two-way ANOVA with Bonferroni's post hoc test. In all analyses, statistical significance was determined at the 5% level ($P < 0.05$). Depicted are mean values $\pm$ s.d. or $\pm$ s.e.m. as indicated in the figure legends. Statistical analysis was performed with Prism5 or Prism6 (GraphPad) or Excel (Microsoft) software.

31. Liang, C. C., Park, A. Y. & Guan, J. L. *In vitro* scratch assay: a convenient and inexpensive method for analysis of cell migration *in vitro*. *Nature Protocols* **2,** 329–333 (2007).
32. Korhonen, H. *et al.* Anaphylactic shock depends on endothelial Gq/G11. *J. Exp. Med.* **206,** 411–420 (2009).
33. Cermak, T. *et al.* Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res.* **39,** e82 (2011).
34. Dittgen, T. *et al.* Lentivirus-based genetic manipulations of cortical neurons and their optical and electrophysiological monitoring in vivo. *Proc. Natl Acad. Sci. USA* **101,** 18206–18211 (2004).
35. Young-Pearse, T. L. *et al.* A critical function for β-amyloid precursor protein in neuronal migration revealed by in utero RNA interference. *J. Neurosci.* **27,** 14459–14469 (2007).
36. Sivaraj, K. K. *et al.* G13 controls angiogenesis through regulation of VEGFR-2 expression. *Dev. Cell* **25,** 427–434 (2013).

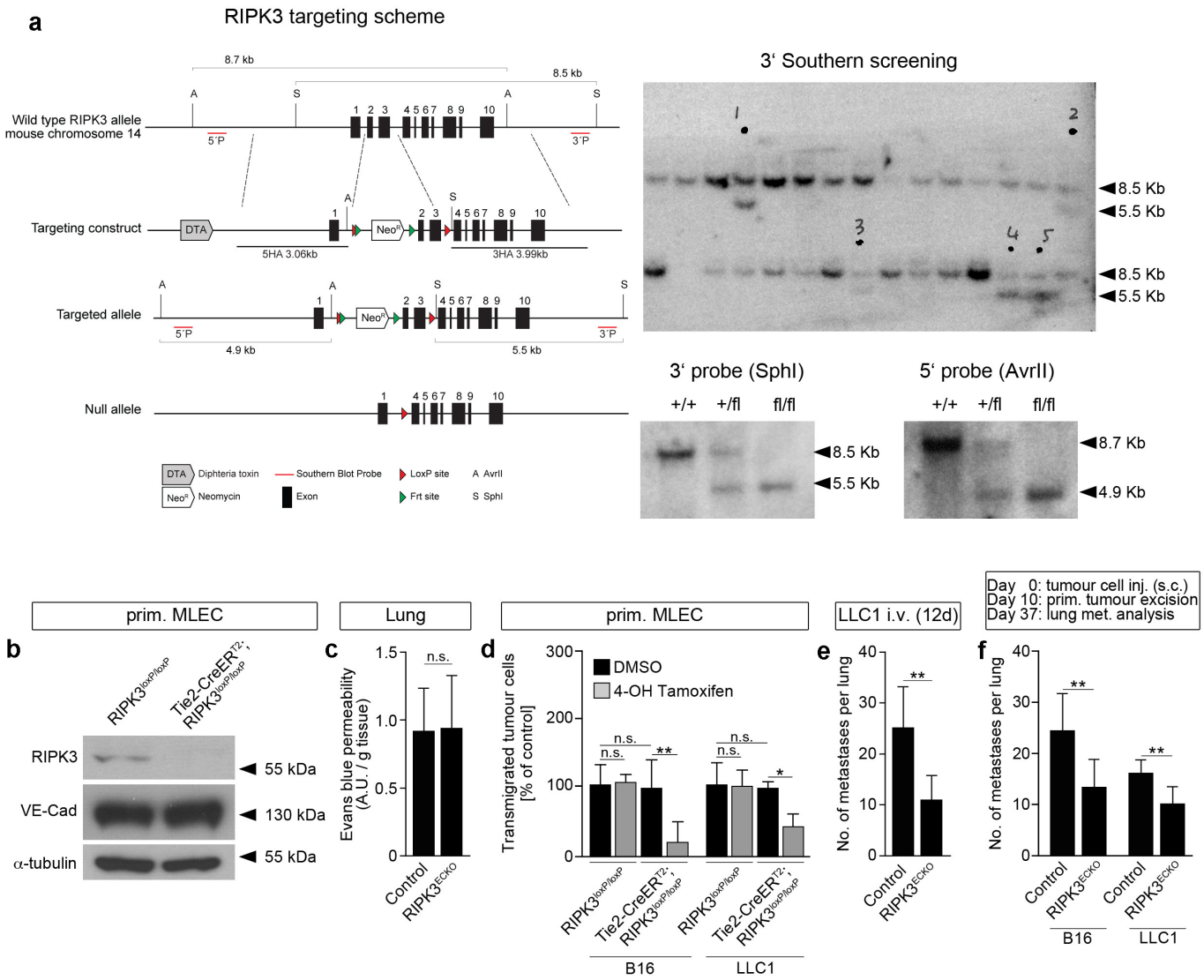


Extended Data Figure 1 | See next page for caption.

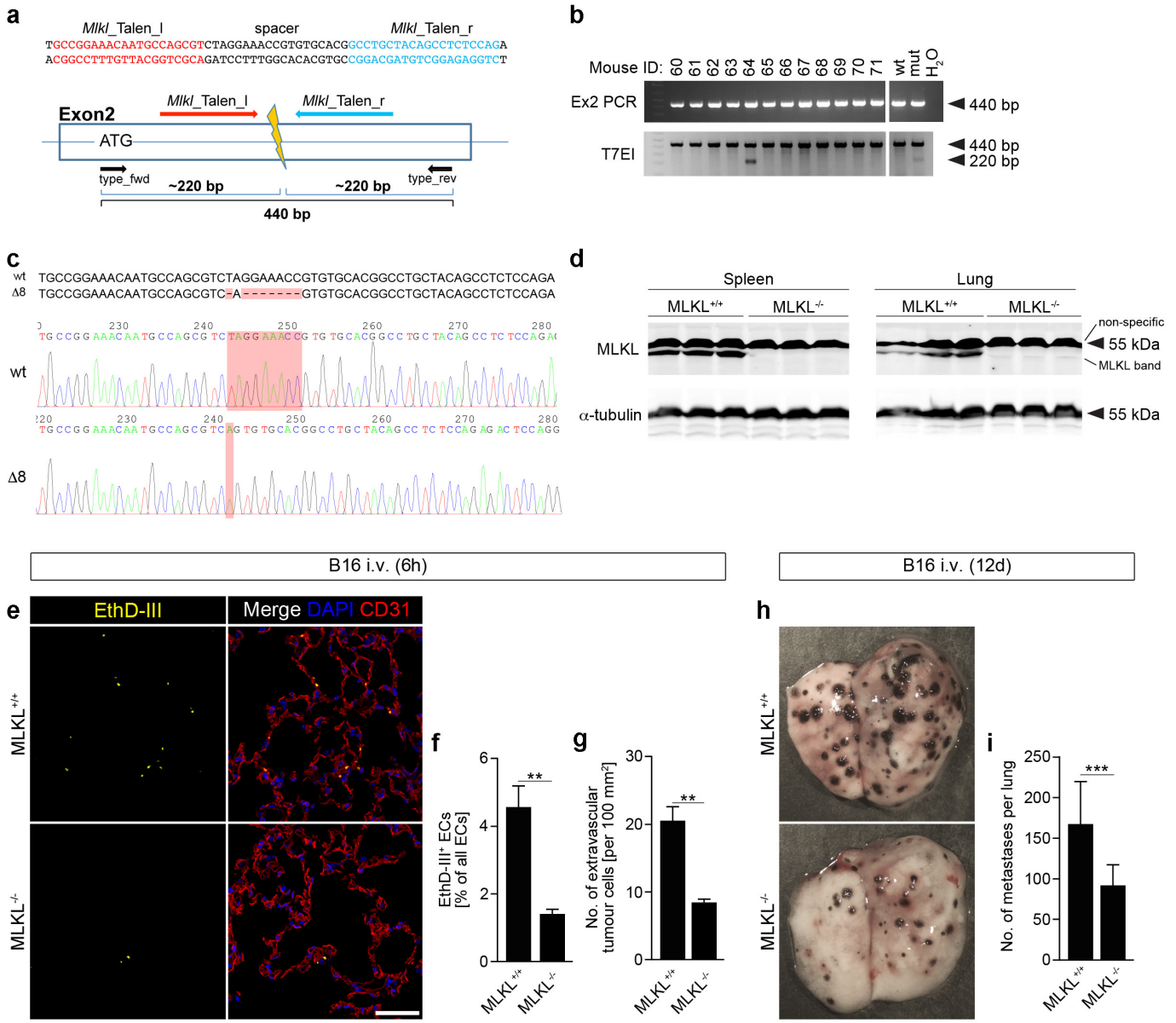
Extended Data Figure 1 | TCs induce necroptosis in ECs. a, Criteria for the discrimination of apoptotic and necrotic EC death. Fluorescent images of HUVECs cultured under conditions to induce apoptosis (TRAIL, Staurosporine) or to induce necrosis (H_2O_2, hypoxia (1% O_2)) and stained with Hoechst 33342 (blue) and the membrane-impermeable dye ethidium homodimer III (EthD-III, red). Asterisks indicate apoptotic ECs (condensed, fragmented nuclei, EthD-III-negative); closed arrowheads indicate necrotic ECs (normal nuclei, EthD-III-positive). Late apoptotic cells are indicated by open arrowheads (condensed/fragmented nuclei, EthD-III-positive); scale bar, 5 μm. **b**, **c**, No annexin-V-positive cells were detected in HUVECs cultured in the presence of TCs (MDA-MB-231). Stimulation with TNF-α served as positive control; scale bar, 20 μm. **d**, Fluorescent images of L929 cells stimulated with TNF-α to induce programmed necrosis (necroptosis, arrowheads). This effect was reversed when cells were additionally cultured with the RIPK1 inhibitor necrostatin-1 (Nec-1); scale bar, 10 μm. **e**, Effect of freshly isolated PBMCs on EC necrosis either directly or in the presence of MDA-MB-231 TCs. PBMCs contained 20 times the amount of platelets (that is, 3×10^4, 9×10^4 or 3×10^5 platelets). **f**–**h**, Quantification of necrosis in HMVEC-L (**f**), in freshly isolated primary mouse lung ECs (prim. MLEC) (**g**) or in HUVEC (**h**) cultured in the presence of different human and mouse TCs (TCs) and at different concentrations as indicated. **i**, Representative confocal images of HUVEC cultured in the absence of TCs (-TC) or presence of TCs (+TC, green) and stained as indicated; scale bar, 5 μm. Quantification of EthD-III- and phospho (p)-MLKL-double-positive ECs (more than 50 EthDIII-positive cells were analysed). **j**, **k**, Analysis of knockdown efficiencies in HUVEC by western blot for RIPK3 and caspase-8 (**j**) or by quantitative RT–PCR for MLKL (**k**). α-Tubulin served as loading control in **j** and relative mRNA expression levels normalized to GAPDH and to the level detected in scramble siRNA-treated samples (siCTRL) are shown in **k**. Shown are representative data of two (**c**, **g**) or three (**e**, **f**, **h**) independent experiments with mean values ± s.e.m. from biological sextuplicates ($n = 6$). $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; n.s., not significant. One-way ANOVA and Bonferroni's post hoc test (**c**, **e**, **g**, **h**) or unpaired, two-tailed Student's *t*-test (**f**). For gel source data see Supplementary Fig. 1.



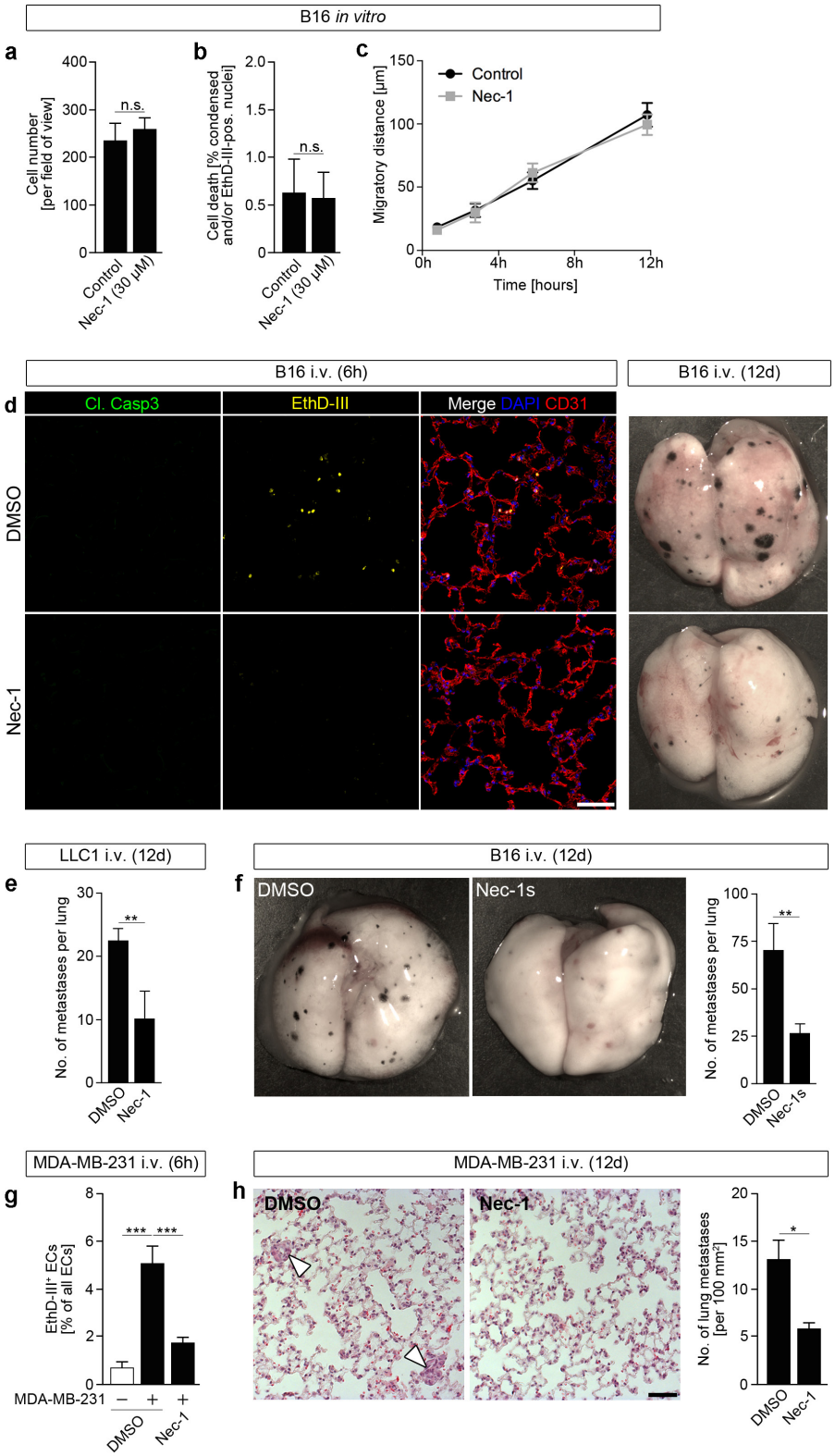
Extended Data Figure 2 | TCs induce necroptosis in ECs *in vivo*. **a**, Quantitative evaluation of cleaved caspase-3- and EthD-III-positive cells in lungs of C57BL/6 WT animals at the indicated time points after i.v. injection of B16 TCs on the basis of the analysis of confocal microscopy images as shown in main Fig. 2a, b. TCs were injected at the same time and lung isolation occurred at the indicated time points. Injection of PBS served as control. **b**, Quantification of EthD-III-positive ECs in lungs of WT animals 6 h after i.v. injection of B16 or LLC1 TCs. **c**–**e**, Representative confocal images of lung sections taken 6 h after i.v. injection of B16 or LLC1 TCs (**c**, **d**) or injection of equal amounts of fluorescently labelled 15 μm microspheres (**e**) into WT animals and stained for the indicated markers. Arrowheads in **c** indicate EthD-III-positive cells. Isolated lungs from animals cultured *ex vivo* in the presence of staurosporine served as positive controls in **d**. Scale bar, 20 μm. Shown are representative data of two (**a**) or three (**b**) independent experiments with mean values ± s.e.m. from $n = 3$ animals per time point (**a**) or $n = 6$ animals per group (**b**). $^{**}P < 0.01$; $^{***}P < 0.001$. One-way ANOVA and Bonferroni's post hoc test.



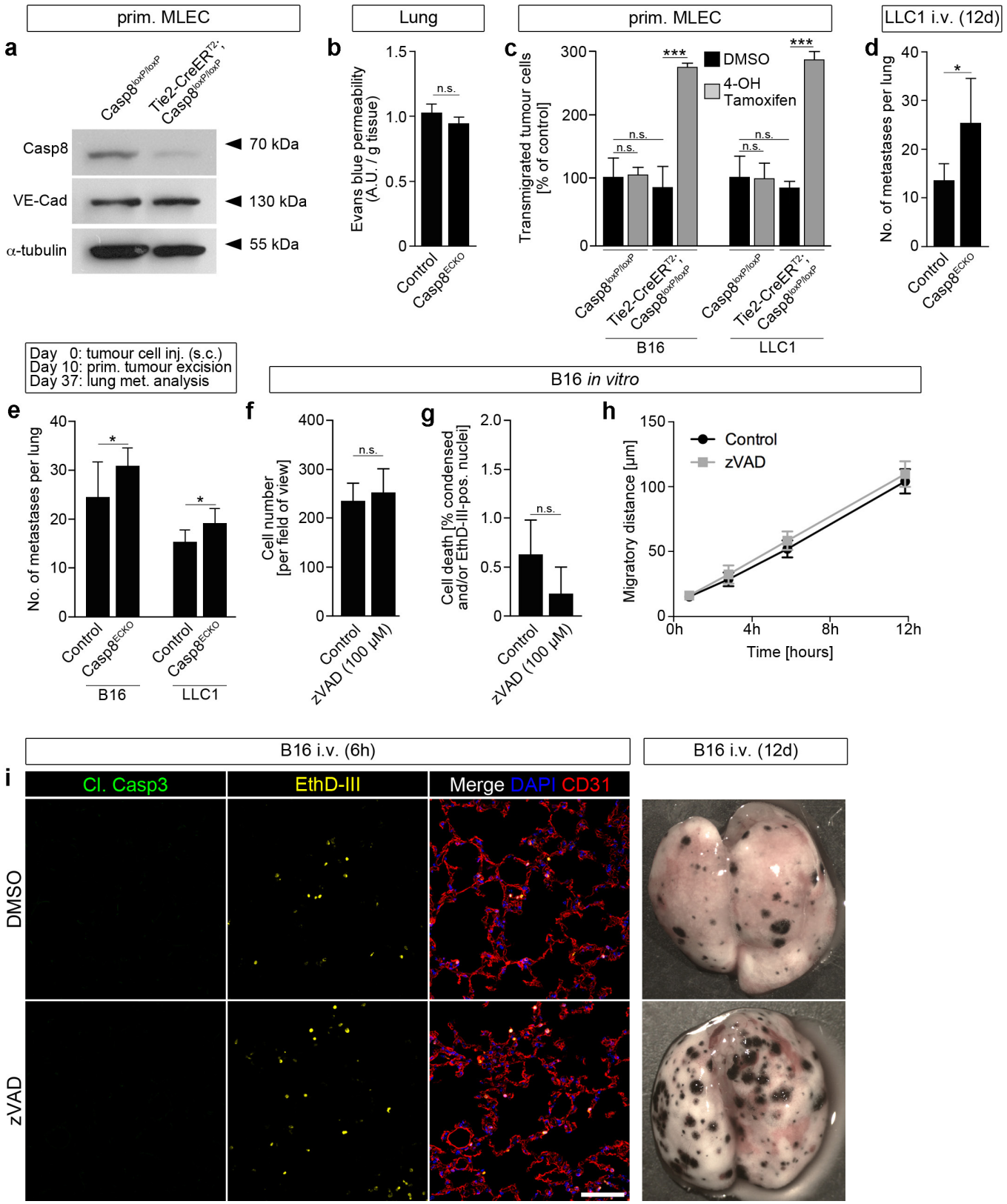
Extended Data Figure 3 | Targeting strategy for the generation of mice with *loxP*-flanked (floxed (fl)) *Ripk3* allele. a, Targeting scheme for the generation of floxed RIPK3 including 3′ Southern screening for the identification of positive ES cell clones and 3′ and 5′ Southern blot confirmation of heterozygous and homozygous floxed *Ripk3* alleles, respectively. **b**, Western blot analysis of RIPK3 and VE-cadherin (VE-Cad) in primary ECs isolated from lungs (MLEC) of tamoxifen-treated Tie2-CreER^{T2};RIPK3$^{loxP/loxP}$ animals (RIPK3^{ECKO}). Cre-negative littermates served as control. α-tubulin served as loading control. **c**, Quantification of Evans blue permeability in the lungs of RIPK3^{ECKO} animals. **d**, Quantification of transmigrated B16 or LLC1 TCs over a layer of DMSO- or 4-OH-tamoxifen-treated primary MLEC isolated from uninduced Tie2-CreER^{T2};RIPK3$^{loxP/loxP}$ animals or Cre-negative control littermates. **e**, **f**, Quantification of lung metastases 12 d after i.v. injection (**e**) or 27 d after excision of a primary tumour induced by s.c. injection (**f**) of B16 or LLC1 TCs into RIPK3^{ECKO} animals. Cre-negative littermates served as control. No significant differences in primary tumour growth were observed (data not shown). Shown are representative data of three (**c**–**e**) or two (**f**) independent experiments with mean values ± s.d. from $n = 3$ (**c**) or $n = 6$ (**e**) animals per group or from $n = 11$ (B16, control), $n = 6$ (B16, RIPK3^{ECKO}), $n = 8$ (LLC1, control) and $n = 6$ (LLC1, RIPK3^{ECKO}) animals (**f**) or $n = 6$ wells per condition (**d**). $^{*}P < 0.05$; $^{**}P < 0.01$; n.s., not significant. Unpaired, two-tailed Student's *t*-test (**c**, **e**, **f**) or one-way ANOVA and Bonferroni's post hoc test (**d**). For gel source data see Supplementary Fig. 1.



Extended Data Figure 4 | Reduced TC-induced endothelial necroptosis and metastasis in MLKL$^{-/-}$ mice. a, A TALEN pair targeting the indicated sequences in the second exon of the mouse *Mlkl* gene was cloned and transcribed into mRNA. **b**, Analysis of mice born after mRNA injection into fertilized oocyte injection by PCR for exon 2 (top) and subsequent sequence specific endonuclease assay (T7EI, bottom) for the detection of mutant alleles. DNA from C57BL/6 (WT) or TALEN-transfected cells (mut) served as control. **c**, The mutant allele of mouse 64 was further analysed using Sanger sequencing revealing a 8 bp deletion (Δ8), predicted to generate a premature stop codon. **d**, Spleen and lung extracts of three B6D2F1 MLKL$^{+/+}$ WT and three homozygous MLKL mutant mice (MLKL$^{-/-}$) were probed for Mlkl protein expression. Polyclonal MLKL antibody detected Mlkl protein only in WT extracts but also an additional non-specific band of similar size. **e**–**i**, B6D2F1 MLKL$^{+/+}$ WT or MLKL$^{-/-}$ animals were injected i.v. with B16 TCs and lungs were analysed after 6 h for pulmonary EthD-III-positive ECs (**f**) and extravascular TCs (**g**) or after 12 d for lung metastases (**i**). Representative confocal images of lung sections and images of whole lungs are shown in (**e**, **h**); scale bar, 50 μm. Shown are representative data of two independent experiments with mean values ± s.e.m. (**f**, **g**) or ± s.d. (**i**) from $n = 3$ animals per condition (**f**, **g**) or $n = 8$ (MLKL$^{+/+}$) and $n = 11$ (MLKL$^{-/-}$) animals (**i**). $^{**}P < 0.01$; $^{***}P < 0.001$. Unpaired, two-tailed Student's *t*-test. For gel source data see Supplementary Figs 1 and 2.

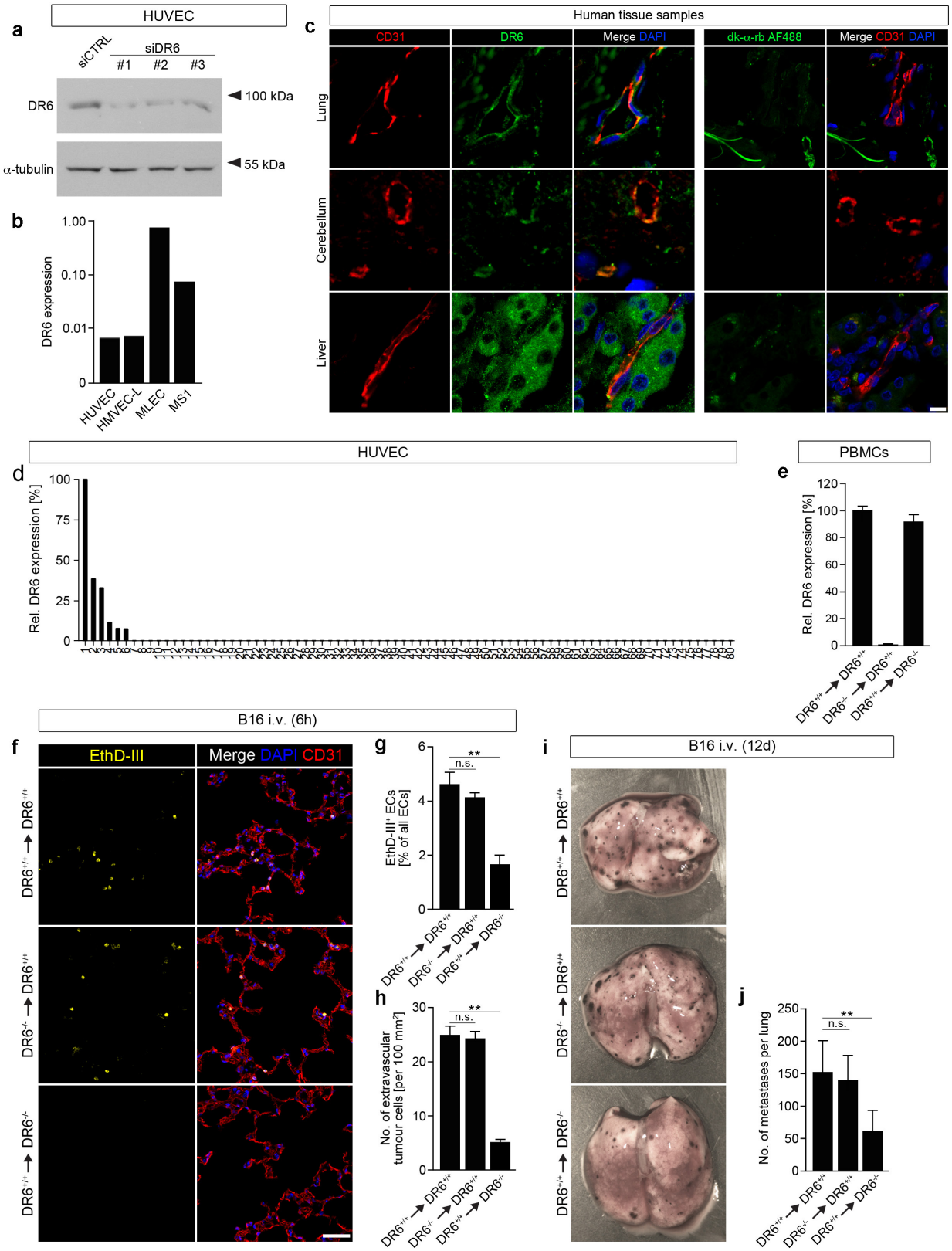


Extended Data Figure 5 | Effect of Nec-1 and Nec-1s treatment on metastasis formation. **a**–**c**, Effect of Nec-1 (30 μM) on B16 TC proliferation (**a**), viability (**b**) and migration (**c**). **d**, Representative confocal images of lung sections taken 6 h and images of whole lungs taken 12 d after i.v. injection of B16 TCs into WT animals treated with DMSO (control) or Nec-1; scale bar, 50 μm. **e**, **f**, Lung metastases 12 d after i.v. injection of LLC1 TCs into WT animals treated with Nec-1 (**e**) or B16 TCs injected into WT animals treated with stable Nec-1s (**f**). DMSO served as control. **g**, **h**, Human MDA-MB-231 TCs were injected i.v. into Nec-1 treated immunodeficient SCID mice and pulmonary EthD-III-positive endothelial and lung metastases were analysed after 6 h and 12d, respectively. Arrowheads in the H&E-stained lung sections (**h**) indicate metastases; scale bar, 50 μm. In all metastasis experiments, animals were treated with Nec-1 or Nec-1s shortly before and at 3 h after TC injection (plus at 6 h for the 12 d experiment). Shown are representative data of three (**a**–**c**, **e**, **g**) or two (**f**, **h**) independent experiments with mean values ± s.e.m. (**a**, **b**, **g**) or ± s.d. (**c**, **e**, **f**, **h**) from biological sextuplicates ($n = 6$) (**a**, **b**), triplicates ($n = 3$) (**c**) or from $n = 3$ (**g**) or $n = 6$ animals per condition (**e**, **f**, **h**). $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; n.s., not significant. Unpaired, two-tailed Student's *t*-test (**a**–**c**, **e**, **f**, **h**) or one-way ANOVA and Bonferroni's post hoc test (**g**).



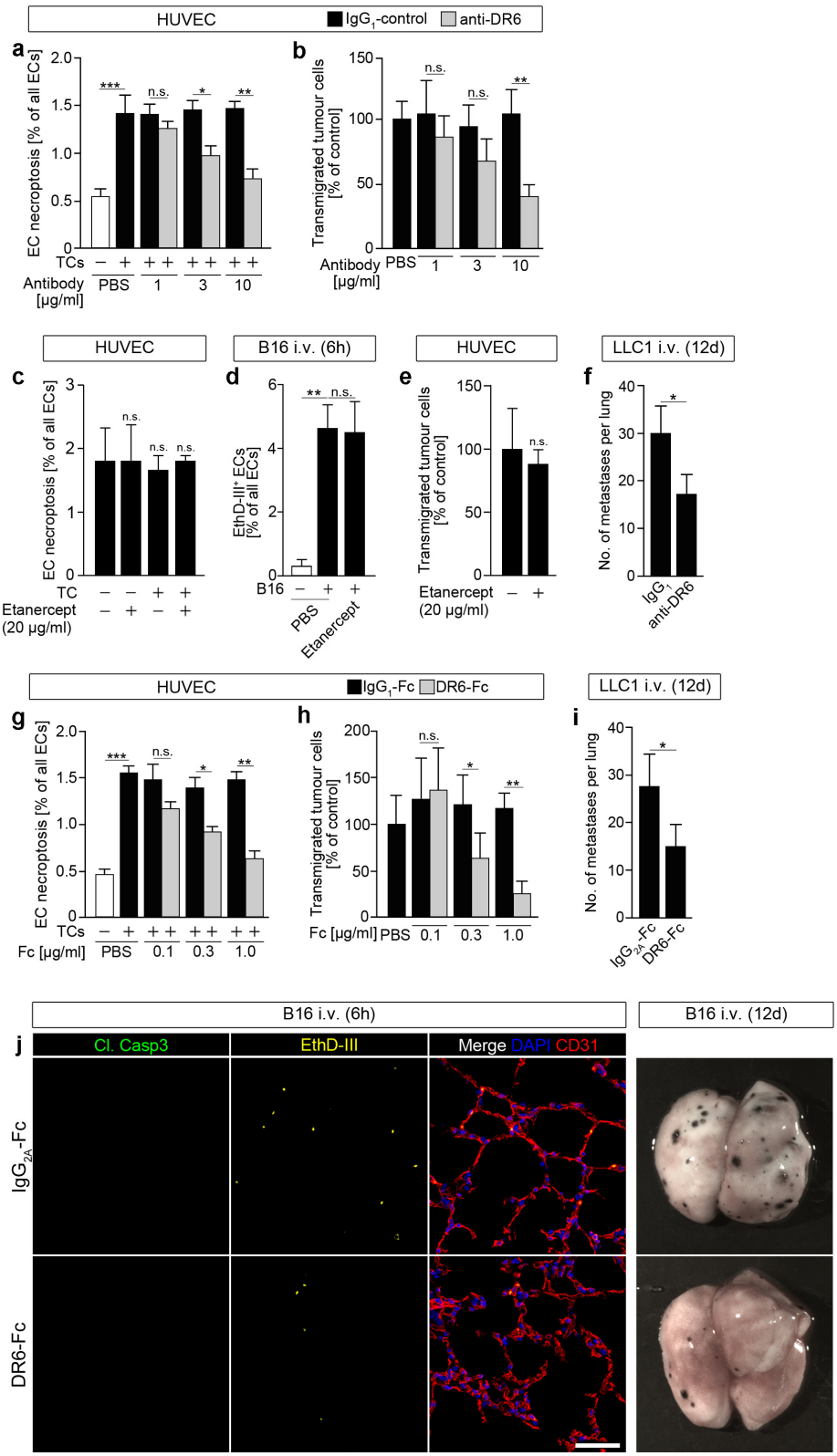
Extended Data Figure 6 | See next page for caption.

Extended Data Figure 6 | Effects of zVAD treatment and EC-specific loss of caspase–8 on metastasis formation. a, Western blot analysis of uncleaved caspase-8 (Casp8) and VE-cadherin (VE-Cad) in primary MLECs of tamoxifen-treated Tie2-CreER^{T2};Casp8$^{loxP/loxP}$ animals (Casp8^{ECKO}). **b**, Quantification of Evans blue permeability in the lungs of Casp8^{ECKO} animals. **c**, Quantification of transmigrated B16 or LLC1 TCs over a layer of DMSO- or 4-OH-tamoxifen-treated primary MLEC from uninduced Tie2-CreER^{T2};Casp8$^{loxP/loxP}$ animals or Cre-negative control littermates. **d**, **e**, Quantification of lung metastases 12 d after i.v. injection (**d**) or 27 d after excision of a primary tumour induced by s.c. injection (**e**) of B16 or LLC1 TCs into Casp8^{ECKO} animals. Cre-negative littermates served as control. No significant differences in primary tumour growth were observed (data not shown). **f**–**h**, Effect of z-VAD-fmk (zVAD, 100 μM) on B16 TC proliferation (**f**), viability (**g**) and migration (**h**). **i**, Representative confocal images of lung sections taken 6 h and images of whole lungs taken 12 d after i.v. injection of B16 TCs into WT animals treated with DMSO (control) or zVAD shortly before and at 3 h after TC injection (plus at 6 h for the 12 d experiment); scale bar, 50 μm. Shown are representative data of two (**b**, **c**, **e**) or three (**d**, **f**–**h**) independent experiments with mean values ± s.d. (**b**–**e**, **h**) or ± s.e.m. (**f**, **g**) from $n = 3$ (**b**) or $n = 5$ (**d**) animals per group or from $n = 11$ (B16, control), $n = 7$ (B16, Casp8^{ECKO}), $n = 8$ (LLC1, control) and $n = = 6$ (LLC1, Casp8^{ECKO}) animals (**e**) or from biological sextuplicates ($n = 6$) (**c**, **f**, **g**) or triplicates ($n = 3$) (**h**). $^{*}P < 0.05$; $^{***}P < 0.001$; n.s., not significant. Unpaired, two-tailed Student's *t*-test (**b**, **d**–**h**) or one-way ANOVA and Bonferroni's post hoc test (**c**). For gel source data see Supplementary Fig. 1.

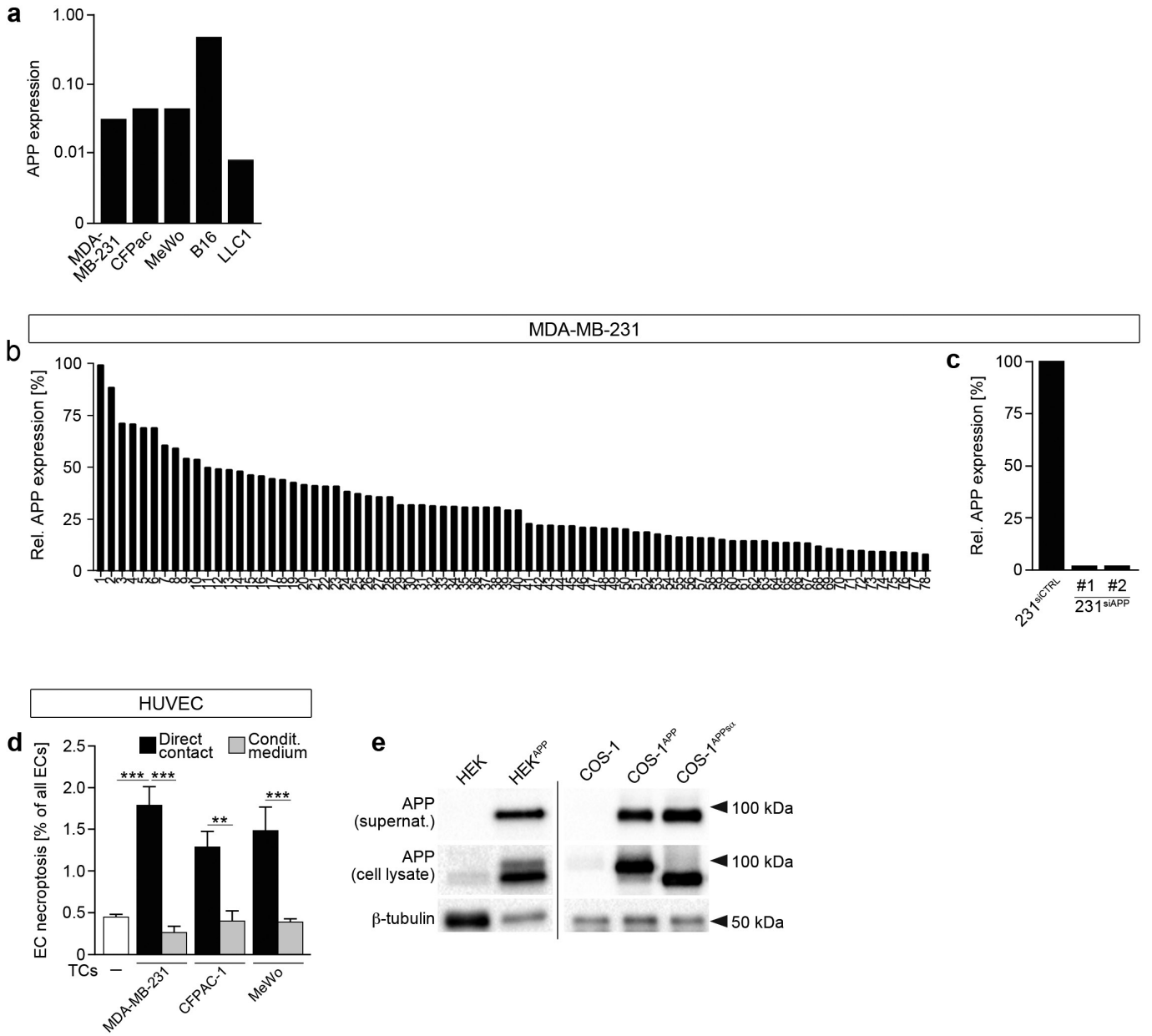


Extended Data Figure 7 | See next page for caption.

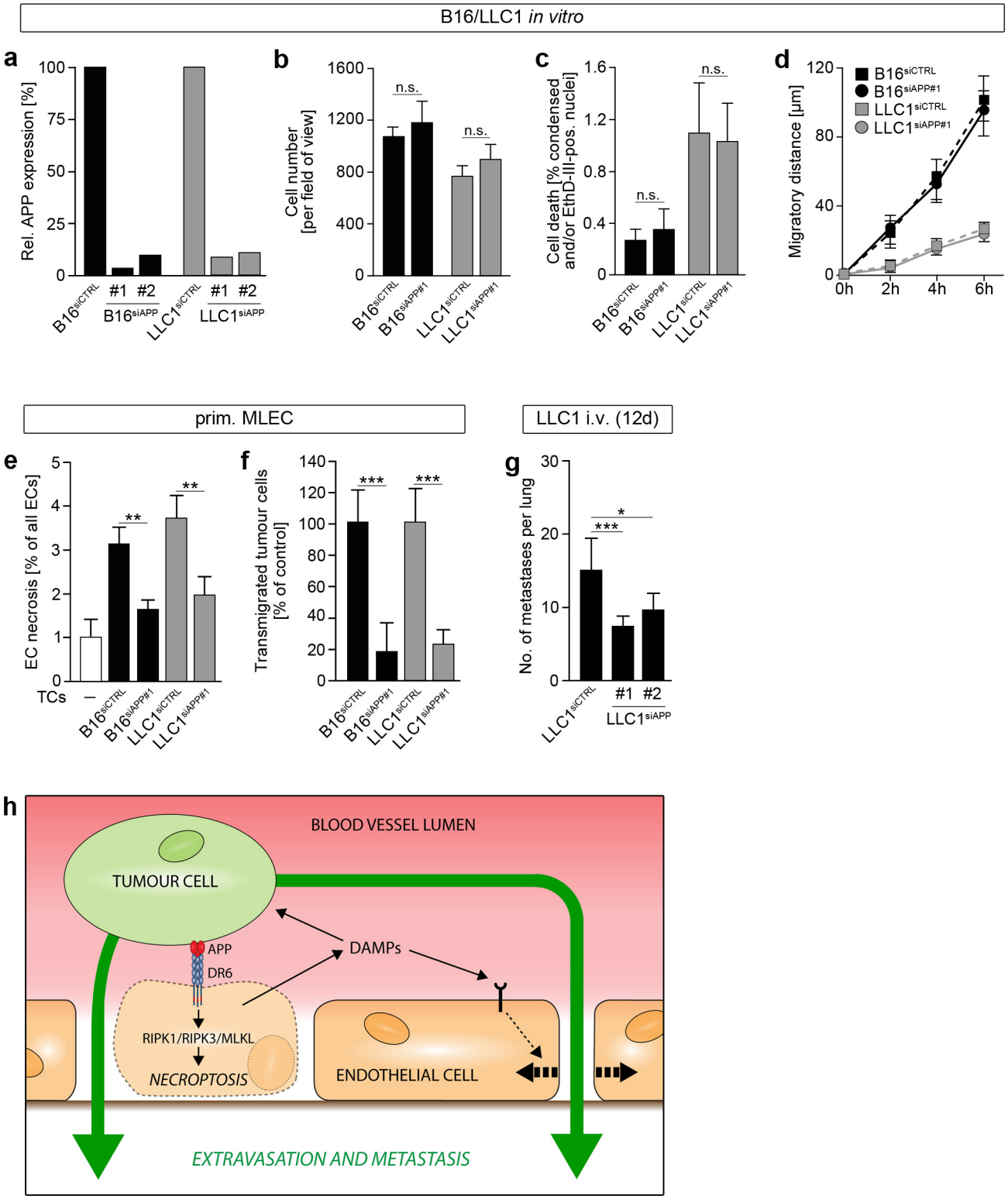
Extended Data Figure 7 | DR6 is expressed in ECs, and DR6 expressed in immune cells is not involved in TC metastasis. a, Western blot analysis of lysates from HUVECs transfected with scrambled siRNA (siCTRL) or different sets of siRNAs directed against mRNA encoding DR6 (siDR6). The antibody detects the 90 kDa glycosylated form of DR6. α-tubulin served as loading control. **b**, Expression levels of DR6 in different human or mouse ECs as determined by quantitative PCR. Shown are relative expression levels normalized to GAPDH levels. **c**, Confocal images of human tissues of the indicated origin stained for CD31 (red), DR6 (green) and cell nuclei (DAPI, blue) (left panel). Control-IgG antibody and donkey anti-rabbit secondary antibody coupled to AF488 served as negative controls (right panels); scale bar, 5 μm. **d**, Analysis of HUVEC single-cell gene expression for DR6 as determined by the $2^{\text{LoDct-ct}}$ method and normalized to the cell with the highest expression level (100%). Each bar represents the gene expression level of one individual cell (data of a total of 80 cells analysed are shown). Single-cell analysis revealed that less than 10% of ECs express DR6. **e**–**j**, Irradiated $DR6^{+/+}$ or $DR6^{-/-}$ animals were reconstituted with bone marrow cells from $DR6^{+/+}$ or $DR6^{-/-}$ donor animals, respectively ($DR6^{+/+} \rightarrow DR6^{+/+}$, $DR6^{-/-} \rightarrow DR6^{+/+}$ or $DR6^{+/+} \rightarrow DR6^{-/-}$) and quantitative PCR analysis of DR6 expression in PBMCs was performed (**e**) or bone marrow chimaeras as indicated were injected i.v. with B16 TCs and lungs were analysed after 6 h for pulmonary EthD-III-positive ECs (**g**) and extravascular TCs (**h**) or after 12 d for lung metastases (**j**). **f**, **i**, Representative confocal images of lung sections stained for the indicated markers (**f**) and images of whole lungs (**i**); scale bar, 50 μm. Shown are representative data of two independent experiments with mean values ± s.d. (**e**, **j**) or ± s.e.m. (**g**, **h**) from $n = 3$ (**e**) or $n = 4$ (**g**, **h**) animals per condition or $n = 6$ ($DR6^{+/+} \rightarrow DR6^{+/+}$), $n = 4$ ($DR6^{-/-} \rightarrow DR6^{+/+}$) and $n = 5$ ($DR6^{+/+} \rightarrow DR6^{-/-}$) animals (**j**). $^{**}P < 0.01$; n.s., not significant. One-way ANOVA and Bonferroni's post hoc test. For gel source data see Supplementary Fig. 1.



Extended Data Figure 8 | Effects of anti-DR6 antibody and DR6-Fc on tumour metastasis. **a**–**i**, Quantification of MDA-MB-231 TC-induced EC necroptosis (**a**, **c**, **g**), transmigrated TCs over an endothelial layer (**b**, **e**, **h**), B16 TC-induced EC necroptosis after i.v. injection into C57BL/6 WT animals (**d**) or lung metastases 12 d after i.v. injection of LLC1 TCs into WT animals (**f**, **i**) upon treatment with an anti-DR6 antibody (5D10) (**a**, **b**, **f**), etanercept (**c**–**e**) or the extracellular domain of DR6 fused to DR6 (DR6-Fc) (**g**–**j**). Animals were treated shortly before and 3 h after TC injection (**d**) as well as 6 h after TC injection (**f**, **i**). PBS, an IgG_1 antibody or the Fc domain of IgG_1 or IgG_{2A} (IgG_1-Fc, IgG_{2A}-Fc) served as controls. **g**, Representative confocal images of lung sections taken 6 h and images of whole lungs taken 12 d after i.v. injection of B16 TCs into WT animals treated with IgG_{2A}-Fc (control) or DR6-Fc shortly before and at 3 h after TC injection (plus at 6 h for the 12 d experiment); scale bar, 50 μm. Shown are representative data of three independent experiments with mean values ± s.e.m. (**a**–**e**, **g**, **h**) or ± s.d. (**f**, **i**) from biological sextuplicates ($n = 6$) (**a**–**c**, **e**, **g**, **h**) or from $n = 3$ (**d**) or $n = 4$ (**f**, **i**) animals per group. $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; n.s., not significant. Two-way ANOVA and Bonferroni's post hoc test (**a**–**d**, **g**, **h**) or unpaired, two-tailed Student's *t*-test (**e**, **f**, **i**).



Extended Data Figure 9 | APP is expressed in different murine and human TCs. a, Expression levels of APP in different human and mouse TCs as determined by quantitative PCR. Shown are relative expression levels normalized to GAPDH levels. **b**, MDA-MB-231 single-cell gene expression analysis for APP as determined by the $2^{\text{LoDct-ct}}$ method and normalized to the cell with the highest expression level (100%). Each bar represents the gene expression level of one individual cell (data of a total of 78 cells analysed are shown). Single-cell analysis revealed that all TCs express APP. **c**, Analysis of knockdown efficiency in MDA-MB-231 TCs using different siRNAs against APP (siAPP). Shown is the relative mRNA expression normalized to GAPDH levels and to the level detected in scramble siRNA-treated samples (siCTRL). **d**, Quantification of EC necroptosis in HUVECs cultured in the presence of MDA-MB-231 TCs (direct contact) or exposed to the supernatant of TC–EC co-cultures after 18 h of culture (condit. medium). **e**, Western blot analysis of cell lysates or conditioned media (supernat.) of parental HEK293T (HEK) cells, HEK cells stably expressing membrane-bound full-length APP_{695} (HEK^{APP}), mock-transfected COS-1 cells (COS-1) or COS-1 cells transiently expressing membrane-bound full-length APP_{695} ($COS\text{-}1^{APP}$) or soluble APPsα ($COS\text{-}1^{APPs\alpha}$). Note that secreted APPsα from COS-1 cells (supernat.) compared with APPsα found in the corresponding cell lysates is glycosylated and thus runs at higher molecular mass. Anti-APP (22C11) was used to detect APP. Membranes were cut to detect β-tubulin as loading control. Shown are representative data of three independent experiments with mean values ± s.e.m. from biological sextuplicates ($n = 6$). $^{**}P < 0.01$; $^{***}P < 0.001$. Two-way ANOVA and Bonferroni's post hoc test. For gel source data see Supplementary Fig. 1.



Extended Data Figure 10 | Effects of loss of APP on TCs and on metastasis formation. a, Analysis of knockdown efficiency in B16 and LLC1 TCs using different siRNAs against mRNA encoding APP (siAPP). Shown is the relative mRNA expression normalized to GAPDH levels and to the level detected in scramble siRNA-treated samples (siCTRL). **b**–**d**, Knockdown of APP in B16 or LLC1 TCs ($B16^{siAPP}$ and $LLC1^{siAPP}$) and evaluation of cell proliferation (**b**), viability (**c**) and migration (**d**). **e**, **f**, Evaluation of APP-deficient TC-induced EC death *in vitro* in C57BL/6 WT primary MLECs (**e**) and the ability of APP-deficient TCs to migrate over an endothelial layer (**f**). **g**, Quantification of lung metastases 12 d after i.v. injection of LLC1 TCs with silenced APP expression ($LLC1^{siAPP}$) into WT animals. Shown are representative data of three independent experiments with mean values ± s.e.m. (**b**, **c**, **e**) or ± s.d. (**d**, **f**, **g**) from biological sextuplicates ($n = 6$) (**b**, **c**, **e**, **f**), triplicates ($n = 3$) (**d**) or from $n = 5$ animals per condition (**g**). $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; n.s., not significant. Unpaired, two-tailed Student's *t*-test (**b**, **c**, **f**) or one-way ANOVA and Bonferroni's post hoc test (**e**, **g**). **h**, Model: TCs induce RIPK1/RIPK3/MLKL-dependent necroptosis in ECs via APP-DR6. TCs then may directly pass through the emerging gap after EC death. Alternatively, or in parallel, damage-associated molecular pattern molecules (DAMPs) released from necroptotic ECs could act on TCs and/or non-necroptotic endothelial as well as other cells to promote TC extravasation and metastasis.

LETTER

doi:10.1038/nature19070

Global profiling of SRP interaction with nascent polypeptides

Daniela Schibich[1,2], Felix Gloge[1,2], Ina Pöhner[3], Patrik Björkholm[4,5,6], Rebecca C. Wade[1,3,7], Gunnar von Heijne[4,5], Bernd Bukau[1,2] & Günter Kramer[1,2]

Signal recognition particle (SRP) is a universally conserved protein–RNA complex that mediates co-translational protein translocation and membrane insertion by targeting translating ribosomes to membrane translocons[1]. The existence of parallel co- and post-translational transport pathways[2], however, raises the question of the cellular substrate pool of SRP and the molecular basis of substrate selection. Here we determine the binding sites of bacterial SRP within the nascent proteome of *Escherichia coli* at amino acid resolution, by sequencing messenger RNA footprints of ribosome–nascent-chain complexes associated with SRP. SRP, on the basis of its strong preference for hydrophobic transmembrane domains (TMDs), constitutes a compartment-specific targeting factor for nascent inner membrane proteins (IMPs) that efficiently excludes signal-sequence-containing precursors of periplasmic and outer membrane proteins. SRP associates with hydrophobic TMDs enriched in consecutive stretches of hydrophobic and bulky aromatic amino acids immediately on their emergence from the ribosomal exit tunnel. By contrast with current models, N-terminal TMDs are frequently skipped and TMDs internal to the polypeptide sequence are selectively recognized. Furthermore, SRP binds several TMDs in many multi-spanning membrane proteins, suggesting cycles of SRP-mediated membrane targeting. SRP-mediated targeting is not accompanied by a transient slowdown of translation and is not influenced by the ribosome-associated chaperone trigger factor (TF), which has a distinct substrate pool and acts at different stages during translation. Overall, our proteome-wide data set of SRP-binding sites reveals the underlying principles of pathway decisions for nascent chains in bacteria, with SRP acting as the dominant triaging factor, sufficient to separate IMPs from substrates of the SecA–SecB post-translational translocation and TF-assisted cytosolic protein folding pathways.

For determination of the nascent substrates of *E. coli* SRP, we used selective ribosome profiling (SeRP)[3] which compares two ribosome profiling data sets generated from the same culture (Extended Data Fig. 1a). The translatome data set comprises the ribosome footprints of all ribosome–nascent-chain complexes (RNCs) and reveals the protein synthesis activity of the cells. The SRP interactome data set comprises the footprints of only SRP-engaged RNCs, purified by immunoprecipitation using SRP-specific antibodies (Extended Data Figs 1b, c). The ratio of both data sets discloses the nascent SRP interactome and provides sequence-resolved interaction profiles of SRP. To enhance the specificity of SeRP, we performed cell lysis and SRP–RNC purifications in the presence of low concentrations of detergents. SeRP without detergent indicated that all relevant features of co-translational SRP action described later were consistently detected (Extended Data Fig. 2).

We identified SRP substrates by ratio-based analysis of translatome and interactome data sets using a threshold of two (SRP interactome/translatome). Almost all SRP substrates identified are IMPs (Figs 1a, b and Extended Data Fig. 1d), suggesting that SRP specifically targets this class of proteins. However, a number of IMPs do not pass the threshold despite their interaction profiles indicating strong transient SRP binding. One example is CopA, which transiently binds SRP but has a ratio of only 0.85 (Extended Data Fig. 1e). We therefore developed a peak detection algorithm that identifies reproducibly detected SRP peaks passing a fivefold threshold. Substrate pools identified by either method overlap largely, but not completely, and together reveal SRP substrates with high confidence (Fig. 1b and Supplementary Table 1).

Among the 2,367 detected nascent *E. coli* proteins are 566 SRP interactors, which according to annotations (Uniprot/Ecocyc) are composed of 488 IMPs, 14 periplasmic/outer membrane proteins, 50 cytoplasmic proteins and 14 proteins with unknown localization. The SRP substrate list includes 87% of all IMPs, 6% of all periplasmic/outer membrane proteins, 3% of all cytoplasmic proteins and 14% of all proteins with unknown localization. Most remaining IMPs do not qualify as SRP substrates either because they do not pass the fivefold threshold (12%) or because of the reduced reproducibility between replicates (86%, Pearson correlation coefficients <0.6), suggesting that these IMPs are SRP substrates as well. These findings indicate a universal function of *E. coli* SRP in IMP targeting, consistent with yeast SRP acting as the predominant targeting factor for TMD-bearing proteins[4]. The five shortest IMPs (<50 amino acids) detected do not interact with SRP (Supplementary Table 1 and Extended Data Fig. 3). Interestingly, targeting of two of these (YbgT (also known as CydX), 37 amino acids; and YbhT (also known as AcrZ), 49 amino acids) is impaired in YidC-depleted cells[5], suggesting that membrane insertion of these and potentially other short IMPs is mediated by YidC but not SRP. This agrees with earlier findings of SRP-independent translocation of small peptides into the endoplasmic reticulum of mammalian cells[6,7].

Our data sets imply that periplasmic and outer membrane proteins, including the assumed SRP substrate DsbA, are efficiently discriminated by SRP. The single DsbA–SRP interaction peak detected in detergent-free SeRP (Extended Data Figs 1f and 2b) is too short to resemble real binding and is not correlated with the DsbA signal sequence. Reminiscent of the yeast SRP substrate pool[8], the largest group of SRP substrates besides IMPs are cytoplasmic proteins, among them the chaperone DnaK (Extended Data Fig. 1g). DnaK functions in protein translocation[9], is partially membrane associated[10], and cooperates with a membrane-bound J-protein cochaperone[11], suggesting that nascent DnaK is partially targeted but not inserted into the membrane. SRP binding to other cytoplasmic proteins may reflect the inherent error rate in SRP substrate recognition. Their cytoplasmic localization implies that these RNCs are rejected during subsequent steps of targeting[12]. We do not detect SRP binding to nascent SecA or σ^{32}, which

[1]Center for Molecular Biology of the University of Heidelberg (ZMBH), DKFZ-ZMBH Alliance, Im Neuenheimer Feld 282, Heidelberg D-69120, Germany. [2]German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, Heidelberg D-69120, Germany. [3]Heidelberg Institute for Theoretical Studies, (HITS gGmbH), Schloss-Wolfsbrunnenweg 35, Heidelberg D-69118, Germany. [4]Department of Biochemistry and Biophysics, Center for Biomembrane Research, Stockholm University, 106 91 Stockholm, Sweden. [5]Science for Life Laboratory, Stockholm University, Box 1031, 171 21 Solna, Sweden. [6]Department of Molecular Evolution, Cell, and Molecular Biology, Uppsala University, 752 36 Uppsala, Sweden. [7]Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Im Neuenheimer Feld 205, Heidelberg D-69120, Germany.

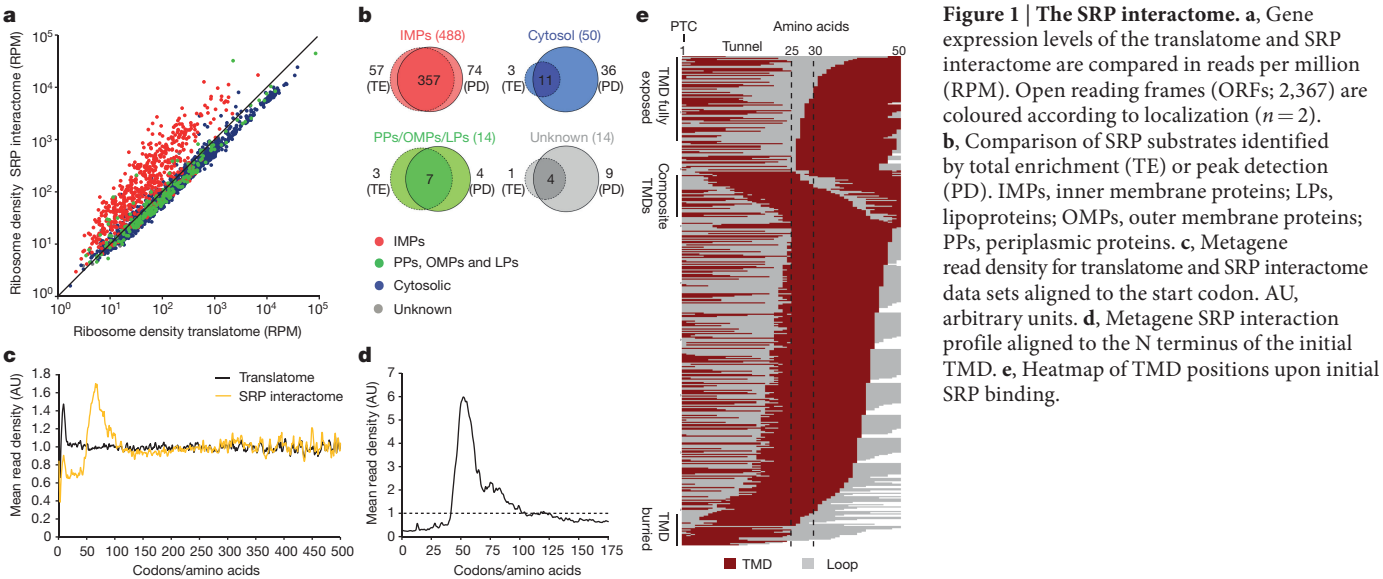


Figure 1 | The SRP interactome. a, Gene expression levels of the translatome and SRP interactome are compared in reads per million (RPM). Open reading frames (ORFs; 2,367) are coloured according to localization ($n = 2$). **b**, Comparison of SRP substrates identified by total enrichment (TE) or peak detection (PD). IMPs, inner membrane proteins; LPs, lipoproteins; OMPs, outer membrane proteins; PPs, periplasmic proteins. **c**, Metagene read density for translatome and SRP interactome data sets aligned to the start codon. AU, arbitrary units. **d**, Metagene SRP interaction profile aligned to the N terminus of the initial TMD. **e**, Heatmap of TMD positions upon initial SRP binding.

was suggested to promote SecA folding[13] and the sensing of proteotoxic stress by σ^{32} (ref. 14).

A proteome-wide metagene analysis of translatome and SRP interactome data shows that, on average, SRP binds RNCs when nascent chains are 50–100 residues long (Fig. 1c). Our results do not support earlier findings that SRP non-discriminately binds all ribosomes early during translation, before the N terminus emerges[15]. To determine the positioning of TMDs facilitating initial SRP engagement, we prepared a ratio metagene interaction profile of all IMPs aligned to the N terminus of the first predicted TMD[16] (Fig. 1d). SRP on average binds an emerging substrate when the N terminus of the first TMD reaches a distance of 42 amino acids from the ribosomal peptidyl transferase centre (PTC); maximal binding is reached at a TMD distance of 55 amino acids from the PTC. Assuming that the 25 C-terminal residues are protected within the ribosome and that SRP senses the nascent chain already in the proximal part of the tunnel, the emergence of about 17 residues of the TMD suffices for SRP recruitment.

Analysing SRP engagement with individual nascent IMPs, we find substrates with remarkably different SRP-binding patterns (Fig. 1e and Extended Data Fig. 4). Some nascent chains are bound when the TMD is distant from the ribosome (Extended Data Fig. 4), while others are bound already upon partial TMD exposure. Another subset exposes two nearby TMDs, suggesting that SRP may bind small segments of one TMD or a composite stretch encompassing neighbouring parts of both TMDs. Finally, we detect some highly sequence-divergent nascent chains that trigger SRP binding before the N-terminal TMD emerges, generally followed by stronger binding once the TMD is fully exposed (Extended Data Figs 4 and 5). The molecular mechanism triggering this early SRP binding remains unclear.

We used our position-resolved information to explore whether SRP binding to the nascent proteome may coincide with a change of average translation speed, as suggested recently[17]. However, the metagene profile of the translatome of 431 SRP substrates aligned to the position of initial SRP binding does not reveal an appreciable peak in the translatome, which would indicate a translation slowdown (Fig. 2a). Thus, considering the translatome as a whole, fidelity and specificity of SRP-mediated targeting in *E. coli* are not generally controlled by variation in translation speed.

We noticed some unexpected SRP binding features to nascent IMPs. First, in 29% of all substrates identified, SRP fails to bind the first TMD but instead binds a more C-terminal TMD. Examples are nascent MsbA, UraA and MetI that bind SRP only after emergence of the second TMD (Fig. 2b and Extended Data Fig. 6a, b). This feature of delayed targeting is independent of the final orientation (N_{in}–C_{out} or N_{out}–C_{in}) of the skipped TMD in the membrane, and is also observed in detergent-free SeRP (Extended Data Fig. 2c, d). TMD skipping is not generally due to false topology predictions of IMPs, since for some IMPs (MsbA, UraA, MetI) crystal structures are available that indicate that the skipped TMDs are membrane embedded. To explore whether skipping is based on intrinsic features of the TMD or its position in the nascent chain, we studied SRP binding to a mutated N-terminal fragment of MsbA encoding both TMDs in switched order and fused to enhanced yellow fluorescent protein (eYFP; MsbA*–eYFP; Fig. 2c). SRP bound MsbA*–eYFP upon emergence of the now N-terminal TMD2, demonstrating that binding is controlled by sequence, but not position of the TMDs. We also analysed SRP binding to purified RNCs that expose TMD1 or TMD2 of MsbA or the DsbA signal sequence on the ribosome surface. Supporting SeRP data, salt-resistant SRP binding to ribosomes is only conferred by TMD2, but not TMD1, of MsbA, or the signal sequence of DsbA (Fig. 2d and Extended Data Fig. 7). A second unexpected feature is that 77% of all IMP substrates are bound multiple times during synthesis, with binding peaks generally correlating with an emerging TMD (Fig. 2b, e and Extended Data Fig. 6c, d).

Both features do not agree with the currently preferred model of co-translational protein targeting, which assumes that SRP engages only the most N-terminal TMD[1]. To account for our observations, we considered the possibility that SRP engages RNCs only once, but because SRP levels *in vivo* are low and perhaps limiting, N-terminal TMDs with low hydrophobicity may sometimes be missed. The delayed SRP binding to more C-terminal TMDs of multi-spanning IMPs will generate multiple SRP-binding peaks in our ensemble measurements. If this model were true, SRP overexpression should facilitate early SRP engagement to the first TMD and reduce binding of SRP to internal TMDs. However, even the overproduction of SRP and FtsY from plasmid did not affect the abundance and amplitude of SRP interaction peaks in nascent IMPs (Extended Data Fig. 6d). We therefore propose that some RNCs may occasionally detach from the translocon and require re-targeting by SRP upon emergence of a downstream TMD. This model agrees with the observations of SRP-dependent re-initiation of translocation in eukaryotes[18] and that length and local hydrophobicity of nascent chains affect the stability of RNC–translocon complexes upon purification from *E. coli*[19].

Mapping of SRP-binding sites allows the identification of nascent chain determinants that mediate binding. We first compared the computed average Gibbs free energy difference of membrane insertion (ΔG_{app})[20] and the Kyte–Doolittle hydrophobicity of N-terminal TMDs that are SRP-skipped or SRP-bound, and of signal sequences. Consistent with earlier work[21], SRP-bound TMDs have higher average

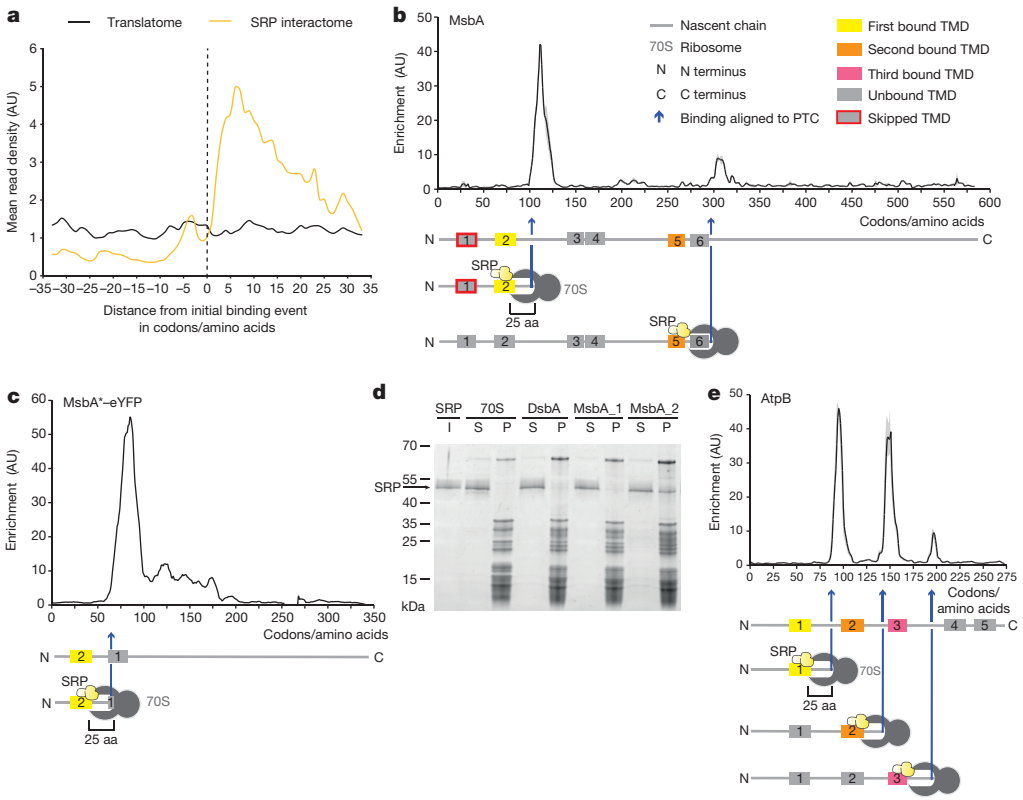


Figure 2 | SRP–RNC interaction. **a**, Metagene translatome and SRP interactome profiles aligned to the position of initial SRP binding (grey dashed line). **b**, Top, SRP interaction profile with nascent MsbA. Light grey shading indicates the variation between two biological replicates. Bottom, cartoon indicating the position of TMDs upon SRP binding. aa, amino acids. **c**, SRP interaction with nascent MsbA*–eYFP ($n = 1$). **d**, SRP interaction with purified RNCs analysed by sucrose cushion centrifugation and SDS–polyacrylamide gel electrophoresis (SDS–PAGE). I, purified SRP; P, ribosomal pellet; S, supernatant. **e**, SRP interaction profile with nascent AtpB. The position of TMDs upon SRP binding is indicated. Shading as in **b**.

hydrophobicity and lower ΔG_{app} than SRP-skipped TMDs (Fig. 3a, b). Signal sequences have the highest ΔG_{app} and lowest hydrophobicity, in particular at their positively charged N termini. Second, we performed a position-resolved analysis of sequence logos[22] to reveal the consensus SRP-binding motif and its distance from the ribosome surface, and compared it with logos for skipped TMDs and signal sequences (Fig. 3c). SRP preferentially binds ribosomes that expose a 12–17-amino-acid-long stretch enriched in hydrophobic residues (Leu, Val, Ile, Phe), starting at a distance of 27 amino acids from the PTC, while skipped TMDs and signal peptides are less hydrophobic. Third, we determined the sequence features of the ribosome-exposed part of the TMDs upon SRP binding (Extended Data Fig. 8 and Supplementary Table 2). SRP prefers binding sites enriched for aliphatic (Met, Leu, Val, Ile) and bulky aromatic (Phe, Trp, Tyr) residues, and a lower content of helix-breaking Pro and Gly residues. Fifty-one per cent of all SRP-bound TMDs have a consecutive stretch of at least four Phe, Ile, Leu or Val residues in

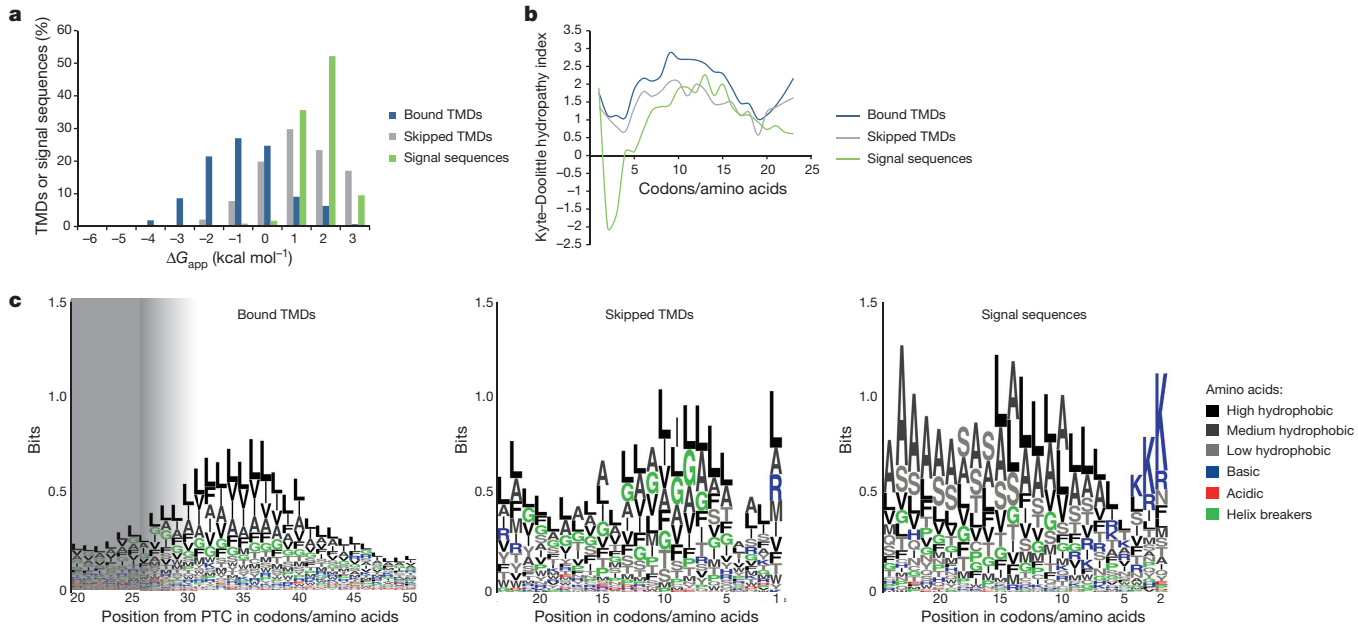


Figure 3 | SRP–nascent-chain-binding properties. **a**, Analysis of computed average free energy differences for membrane insertion, ΔG_{app}. **b**, Position-resolved hydropathies for bound TMDs, skipped TMDs and signal sequences. *x* Axis shows distance from N terminus **c**, Weblogo representations of the amino acid compositions of SRP-bound IMPs upon initial SRP binding, aligned to the C terminus (left; the grey area indicates the ribosomal tunnel), of skipped TMDs aligned to their N terminus (centre), and of signal sequences aligned to second amino acid from the N terminus (right).

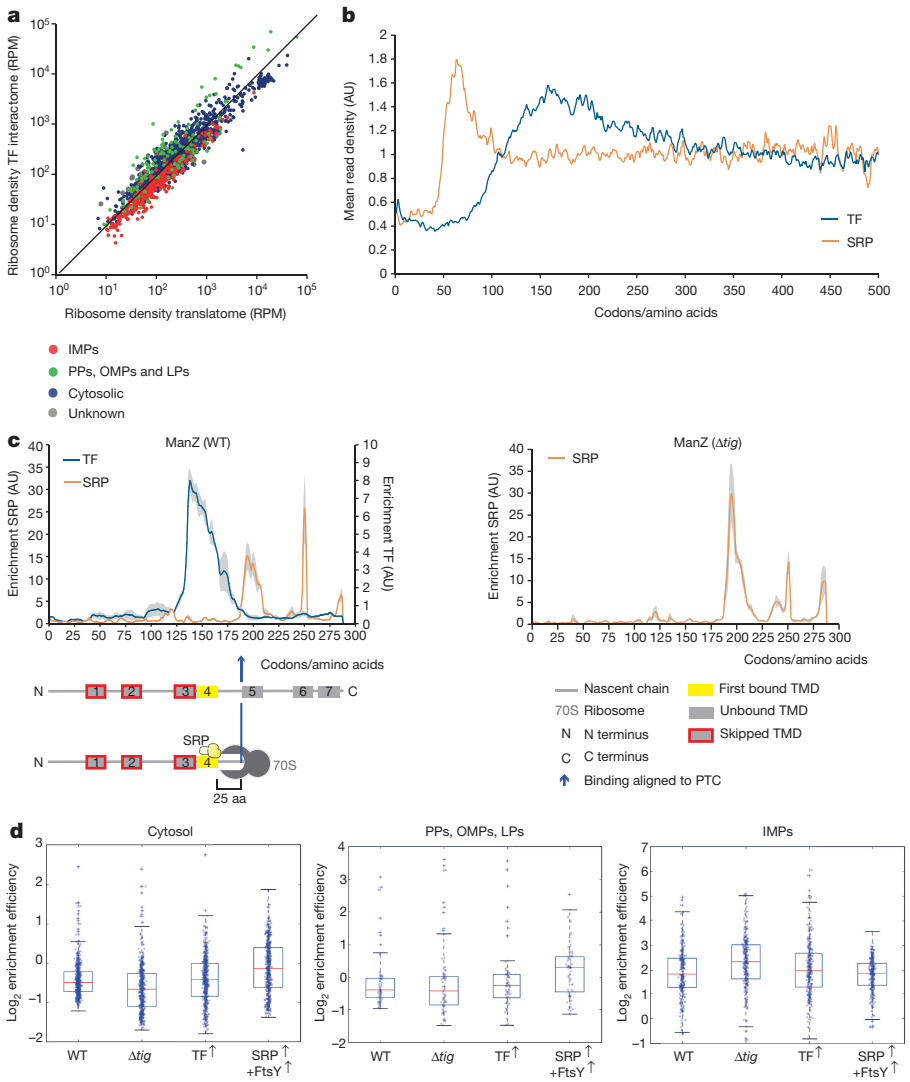


Figure 4 | SRP and TF specificity. a, Gene expression levels in translatome and TF interactome data ($n = 2$). **b**, Metagene interaction profiles of SRP and TF. **c**, Left, TF and SRP interaction profiles with ManZ. Right, SRP interaction profile in cells lacking TF (*Δtig*). Light grey shading indicates the variation between two biological replicates. aa, amino acids; WT, wild type. **d**, Box plot representation of enrichment efficiencies in SRP interactome data sets for proteins of different localizations (cytosol, periplasm, outer membrane, lipoproteins and IMPs) in wild-type cells, cells lacking TF (*Δtig*), cells overexpressing TF (TF↑) and cells overexpressing SRP and FtsY (SRP↑+FtsY↑).

arbitrary order, while only 14% of skipped TMDs and 8% of signal sequences have such a stretch. In agreement with this, the SRP-bound TMD2 of MsbA, but not TMD1, contains a four-residue-long hydrophobic stretch (LVVI) and a lower ΔG_{app} (+1.05 kcal mol^{-1} (TMD1), −1.02 kcal mol^{-1} (TMD2)). In addition to the strong N-terminal enrichment of Lys and their reduced hydrophobicity, signal sequences have fewer aromatic residues and helix breakers than TMDs.

The chaperone TF has been suggested to enhance substrate specificity of SRP *in vivo* and *in vitro*[23,24], implying an important contribution of TF to SRP function. Some studies indicate that SRP and TF compete for ribosome or substrate binding[25], while other studies suggest that they co-exist on ribosomes[23]. To resolve these inconsistencies, we determined the functional importance of TF for SRP substrate binding *in vivo*. SeRP data for both factors show very limited substrate overlap (Figs 1a and 4a). TF avoids binding IMPs and preferentially engages cytoplasmic, periplasmic and outer membrane proteins, and metagene analyses show that SRP and TF engage the bulk of nascent substrates at different time points during translation (Fig. 4b). For the few substrates engaged by both factors, we find strict temporal separation of RNC binding. Owing to the N-terminal position of most TMDs and the generally delayed binding of TF to nascent chains[3], SRP mostly binds before TF (for example, MrcA; Extended Data Fig. 9). One exception is nascent ManZ, which is bound first by TF. The reason for this unusual binding pattern is that SRP skips the three N-terminal TMDs of ManZ and only engages the fourth TMD positioned far from the N terminus (amino acids 147–167; Supplementary Table 1 and Fig. 4c). We also analysed the SRP interactome in cells lacking or overexpressing the *tig* gene encoding TF (Fig. 4d). Contrasting the proposed function of TF as a factor improving SRP specificity[24], we find that the overexpression of *tig* does not affect SRP binding, and rather that *tig* deletion enhances SRP binding to IMPs and decreases binding to cytosolic proteins. Accordingly, the onset of initial SRP binding in ManZ is not changed in the absence of TF (Fig. 4c). SRP binding specificity was relaxed only by transient overexpression of SRP and its receptor FtsY, which triggered more prominent SRP binding to cytoplasmic, periplasmic and outer membrane proteins (Fig. 4d).

Together, our analyses show that SRP, owing to its strong preference for TMDs, acts as the dominant and chaperone-independent router that specifically triages IMPs into the co-translational translocation pathway while periplasmic proteins and outer membrane proteins are exported post-translationally. We speculate that cells may employ the co-translational pathway specifically for IMPs because its slower

translocation kinetics[26,27] alleviates critical steps of membrane insertion, for example, by facilitating helix formation or diffusion of TMDs into the lipid bilayer. In contrast, the faster SecA-dependent post-translational translocation is compatible with folding of periplasmic and outer membrane proteins, and may help to reduce the number of translocons required per cell.

Our data suggest a revised model of co-translational protein translocation, which often initiates on internal TMDs. We speculate that substrates containing N-terminal skipped TMDs may enter the translocon pore by either inserting a loop or additionally engaging SecA for translocating the periplasmic parts of the protein. This model is supported by previous studies suggesting that SecA is involved in co-translational translocation of some IMPs[28].

TF is not involved in substrate selection and acts on different substrate classes and at different time points during translation than SRP, contradicting previous reports[23,24]. This distinguishes TF from the nascent-polypeptide-associated complex NAC of eukaryotes, which, unlike TF, acts as an antagonist of SRP to sharpen SRP specificity in *Caenorhabditis elegans*[29,30]. We conclude that bacterial SRP suffices to triage IMPs to the co-translational pathway.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 30 May; accepted 29 June 2016.
Published online 3 August 2016.

1. Akopian, D., Shen, K., Zhang, X. & Shan, S. O. Signal recognition particle: an essential protein-targeting machine. *Annu. Rev. Biochem.* **82,** 693–721 (2013).
2. Driessen, A. J. & Nouwen, N. Protein translocation across the bacterial cytoplasmic membrane. *Annu. Rev. Biochem.* **77,** 643–667 (2008).
3. Oh, E. *et al.* Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor *in vivo*. *Cell* **147,** 1295–1308 (2011).
4. Ast, T., Cohen, G. & Schuldiner, M. A network of cytosolic factors targets SRP-independent proteins to the endoplasmic reticulum. *Cell* **152,** 1134–1145 (2013).
5. Fontaine, F., Fuchs, R. T. & Storz, G. Membrane localization of small proteins in *Escherichia coli*. *J. Biol. Chem.* **286,** 32464–32474 (2011).
6. Schlenstedt, G., Gudmundsson, G. H., Boman, H. G. & Zimmermann, R. Structural requirements for transport of preprocecropinA and related presecretory proteins into mammalian microsomes. *J. Biol. Chem.* **267,** 24328–24332 (1992).
7. Johnson, N. *et al.* TRC40 can deliver short secretory proteins to the Sec61 translocon. *J. Cell Sci.* **125,** 3612–3620 (2012).
8. del Alamo, M. *et al.* Defining the specificity of cotranslationally acting chaperones by systematic analysis of mRNAs associated with ribosome-nascent chain complexes. *PLoS Biol.* **9,** e1001100 (2011).
9. Castanié-Cornet, M. P., Bruel, N. & Genevaux, P. Chaperone networking facilitates protein targeting to the bacterial cytoplasmic membrane. *Biochim. Biophys. Acta* **1843,** 1442–1456 (2014).
10. Bukau, B., Reilly, P., McCarty, J. & Walker, G. C. Immunogold localization of the DnaK heat shock protein in *Escherichia coli* cells. *J. Gen. Microbiol.* **139,** 95–99 (1993).
11. Clarke, D. J., Jacq, A. & Holland, I. B. A novel DnaJ-like protein in *Escherichia coli* inserts into the cytoplasmic membrane with a type III topology. *Mol. Microbiol.* **20,** 1273–1286 (1996).
12. Zhang, X., Rashid, R., Wang, K. & Shan, S. O. Sequential checkpoints govern substrate selection during cotranslational protein targeting. *Science* **328,** 757–760 (2010).
13. Nakatogawa, H., Murakami, A., Mori, H. & Ito, K. SecM facilitates translocase function of SecA by localizing its biosynthesis. *Genes Dev.* **19,** 436–444 (2005).
14. Lim, B. *et al.* Heat shock transcription factor σ^{32} co-opts the signal recognition particle to regulate protein homeostasis in *E. coli*. *PLoS Biol.* **11,** e1001735 (2013).
15. Bornemann, T., Jöckel, J., Rodnina, M. V. & Wintermeyer, W. Signal sequence-independent membrane targeting of ribosomes containing short nascent peptides within the exit tunnel. *Nature Struct. Mol. Biol.* **15,** 494–499 (2008).
16. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305,** 567–580 (2001).
17. Fluman, N., Navon, S., Bibi, E. & Pilpel, Y. mRNA-programmed translation pauses in the targeting of *E. coli* membrane proteins. *eLife* **3,** e03440 (2014).
18. Kuroiwa, T., Sakaguchi, M., Omura, T. & Mihara, K. Reinitiation of protein translocation across the endoplasmic reticulum membrane for the topogenesis of multispanning membrane proteins. *J. Biol. Chem.* **271,** 6423–6428 (1996).
19. Bischoff, L., Wickles, S., Berninghausen, O., van der Sluis, E. O. & Beckmann, R. Visualization of a polytopic membrane protein during SecY-mediated membrane insertion. *Nature Commun.* **5,** 4103 (2014).
20. Hessa, T. *et al.* Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature* **450,** 1026–1030 (2007).
21. Lee, H. C. & Bernstein, H. D. The targeting pathway of *Escherichia coli* presecretory and integral membrane proteins is specified by the hydrophobicity of the targeting signal. *Proc. Natl Acad. Sci. USA* **98,** 3471–3476 (2001).
22. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14,** 1188–1190 (2004).
23. Hoffmann, A., Bukau, B. & Kramer, G. Structure and function of the molecular chaperone Trigger Factor. *Biochim. Biophys. Acta* **1803,** 650–661 (2010).
24. Ariosa, A., Lee, J. H., Wang, S., Saraogi, I. & Shan, S. O. Regulation by a chaperone improves substrate selectivity during cotranslational protein targeting. *Proc. Natl Acad. Sci. USA* **112,** E3169–E3178 (2015).
25. Bornemann, T., Holtkamp, W. & Wintermeyer, W. Interplay between trigger factor and other protein biogenesis factors on the ribosome. *Nature Commun.* **5,** 4180 (2014).
26. Pugsley, A. P. The complete general secretory pathway in Gram-negative bacteria. *Microbiol. Rev.* **57,** 50–108 (1993).
27. Kadokura, H. & Beckwith, J. Detecting folding intermediates of a protein as it passes through the bacterial translocation channel. *Cell* **138,** 1164–1173 (2009).
28. Sääf, A., Andersson, H., Gafvelin, G. & von Heijne, G. SecA-dependence of the translocation of a large periplasmic loop in the *Escherichia coli* MalF inner membrane protein is a function of sequence context. *Mol. Membr. Biol.* **12,** 209–215 (1995).
29. Gamerdinger, M., Hanebuth, M. A., Frickey, T. & Deuerling, E. The principle of antagonism ensures protein targeting specificity at the endoplasmic reticulum. *Science* **348,** 201–207 (2015).
30. Kramer, G., Guilbride, D. L. & Bukau, B. Finding nascent proteins the right home. *Science* **348,** 182–183 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank members of the Bukau laboratory for valuable contributions; C. Gläßer for support with data analysis; and H. Bernstein for providing plasmid pHQ4. Sequencing was done at the Genomics & Proteomics Core (DKFZ) facility. I.P. and R.C.W. acknowledge support from the Klaus Tschira Foundation. This work was supported by research grants from the Deutsche Forschungsgemeinschaft (SFB638 and FOR1805) to G.K. and B.B., a Human Frontier Science Program grant to B.B. and a grant from the Swedish Research Council to G.v.H.

Author Contributions B.B. and G.K. conceived the study. D.S., B.B. and G.K. designed the experiments. D.S. and F.G. performed the experiments. D.S., I.P., R.C.W., P.B., G.vH., B.B. and G.K. analysed the data. B.B. and G.K. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Author Information Data have been deposited in the Figshare database and are accessible from https://dx.doi.org/10.6084/m9.figshare.2058051. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to G.K. (g.kramer@zmbh.uni-heidelberg.de) or B.B. (bukau@zmbh.uni-heidelberg.de).

Reviewer Information *Nature* thanks N. Stern-Ginossar and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Purification of SRP–RNCs for SeRP. *E. coli* cells (MC4100; ref. 31) were grown in 200 ml EZ Rich Defined Medium (EZ-RDM, Teknova), a rich supplemented MOPS defined media at 37 °C to an $OD_{600\,nm}$ of 0.45. For growth of strains encoding avi-tagged TF, 100 μg/ml ampicillin and 40 μg/ml D-biotin was added. For transient overexpression of *tig*, cells transformed with pTrc-TF[32] were grown in the presence of 250 μM IPTG for one duplication period. For SRP/FtsY overexpression, cells were transformed with the plasmid pHQ4 (ref. 21), facilitating the overexpression of *ffs*, *ffh* and *ftsY*. Growth, harvest and lysis of cells for SeRP were essentially done as described[33]. Briefly, cells were harvested by rapid filtration (pore size 0.2 μm) and frozen in liquid nitrogen. Lysis buffer (50 mM HEPES-KOH 7.5, 100 mM NaCl, 10 mM $MgCl_2$, 5 mM $CaCl_2$, 1 mM chloramphenicol, 1 mM PMSF, 0.4% Triton-X100, 0.1% NP-40, 50 mM octyl-β-D-glucopyranoside) was frozen in liquid nitrogen. Frozen cells were mixed with 500 μl frozen lysis buffer supplemented with 1.3 μl DNaseI and lysed by mixer milling (Retsch MM400, 10 ml jars, 2 min 30 Hz). To the pulverized cells 500 μl cold lysis buffer was added. The RNA concentration was determined and polysomes were digested using MNase (150 U/1 $A_{260\,nm}$) for 5 min at 25 °C. The reaction was terminated by addition of 6 mM EGTA and chilling on ice. Cell debris was removed by 5 min centrifugation (3,000 r.p.m., 4 °C). Monosomes were purified by sucrose cushion centrifugation (lysis buffer lacking $CaCl_2$ and supplemented with 30% sucrose) for 90 min, 75,000 r.p.m., at 4 °C (Beckmann AT2 S120 rotor). Pelleted ribosomes were washed once and resuspended in lysis buffer lacking $CaCl_2$. Detergent-free SeRP was performed in the absence of Triton X-100, NP-40 and octyl β-D-glucopyranoside.

Selective purification of factor bound RNCs. *Immunoprecipitation of SRP–RNCs.* Per 200 ml filtered cells, 2.5 ml Dynabeads (Life Technologies) were used. Beads were washed three times with 5 ml PBS and incubated with 100 μl rabbit anti-SRP antibody generated in our laboratory for 10 min at room temperature under constant shaking. The beads were washed three times with 5 ml buffer (TBS, 1 mM chloramphenicol, 10 mM $MgCl_2$, 0.1% Tween-20). Monosomes were incubated with the affinity beads for 15 min at 4 °C under constant shaking. The matrix was quickly washed three times with cold wash buffer and RNA was extracted by phenol-chloroform extraction. For detergent-free SeRP, Tween-20 was omitted.

Affinity purification of TF-RNCs. To equilibrate the Strep-Tactin sepharose, a 50% slurry was washed twice with 700 μl lysis buffer (50 mM HEPES-KOH 7.5, 100 mM NaCl, 10 mM $MgCl_2$, 1 mM chloramphenicol, 1 mM PMSF, 0.4% Triton-X, 0.1% NP-40, 50 mM octyl β-D-glucopyranoside) for 5 min at 4 °C. 2.5 ml of matrix was used per 200 ml of cells. Monosomes were prepared as described before, omitting the chemical crosslinking reaction and incubated with the affinity matrix for 1 h at 4 °C under constant shaking. The matrix was washed three times with cold lysis buffer (without $CaCl_2$). For TEV cleavage, the matrix was incubated for 1 h in cleavage buffer (50 mM Tris-HCl pH 7.0, 200 mM NaCl, 10 mM $MgCl_2$, 1 mM chloramphenicol). Elution was performed in three consecutive steps. In each step 100 μl of cleavage buffer supplemented with nucleic acid-free TEV protease was added and incubated for 30 min at room temperature. The RNA of the pooled elution fractions was extracted by phenol-chloroform extraction. Detergent-free SeRP was performed in the absence of Triton-X, NP-40 and octyl β-D-glucopyranoside.

Deep sequencing library. Ribosomal footprints were isolated and prepared for deep sequencing as previously described[33].

Data analysis. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment. Sequencing reads were processed as described previously[33]. Further analyses were done using customized python scripts. Additional information on the code will be provided upon request.

Peak detection analysis. SRP substrate identification was based on two methods, first on the ratio-based enrichment passing a threshold of 2.0 and second via peak detection. For peak detection the ratio over a window of 11 nucleotides (nt) of RPM-normalized interactome and translatome data of two replicates was built. Reproducibility of SRP interaction profiles was evaluated in the quality control step by Pearson correlation analysis. If a threshold of 0.6 was passed, genes were processed further. A continuous wavelet transformation was used to detect SRP-binding peaks (min_peak_width = 6 nt, max_peak_width = 90 nt, min_length = 9, gap_thresh = 50). SRP-binding peaks needed to pass a threshold of 5.0 in two replicates.

TMD prediction. To predict transmembrane domains in proteins, we used the ΔG prediction server, http://dgpred.cbr.su.se (ref. 20). The advantage of this method is that is uses a biological membrane insertion scale that has been developed from experimental data[34]. Each hydrophobic segment receives a score that is 3 or less; this score represents the insertion free energy difference (in kcal/mol) between the translocon channel and the bilayer. Each hydrophobic segment predicted by this server was seen as a potential transmembrane domain in the analysis done later with this data. The server script was run with default server settings (Helix min length: 19; Helix max length: 23; Length correction: ON).

Identification of peptide properties. The identification of peptide properties associated with SRP binding was performed as follows. SRP-bound TMDs were compared to those TMDs skipped by SRP using a reference pool of typical SRP substrates. The number of residues contributing to a TMD in the recognition range between residues 26–50 was evaluated and only those 299 substrate peptides with a TMD exposure of at least 17 residues length were kept for further analysis. The TMD region and the additional C- or N-terminal flanking regions within the recognition site were analysed separately. The average number of residues in a TMD and in the flanking regions for the pool of 299 substrates was determined, thereby resulting in average lengths of 17 residues for a TMD, 3 residues for a C-terminal flanking region and 4 residues for an N-terminal flanking region. For each skipped TMD, a C-terminal flanking region of 3 residues and a TMD of 17 residues (counting from the C terminus, 77 sequences in total) were collected for comparison with bound TMD and C-terminal flanking regions, as well as a TMD of 17 residues (counting from the N-terminal end, 87 sequences in total) and an N-terminal flanking region of 4 residues for comparison with bound TMD and N-terminal flanking regions. If the flanking regions were found to overlap with the following or previous TMDs, the TMD with its flanking region was skipped.

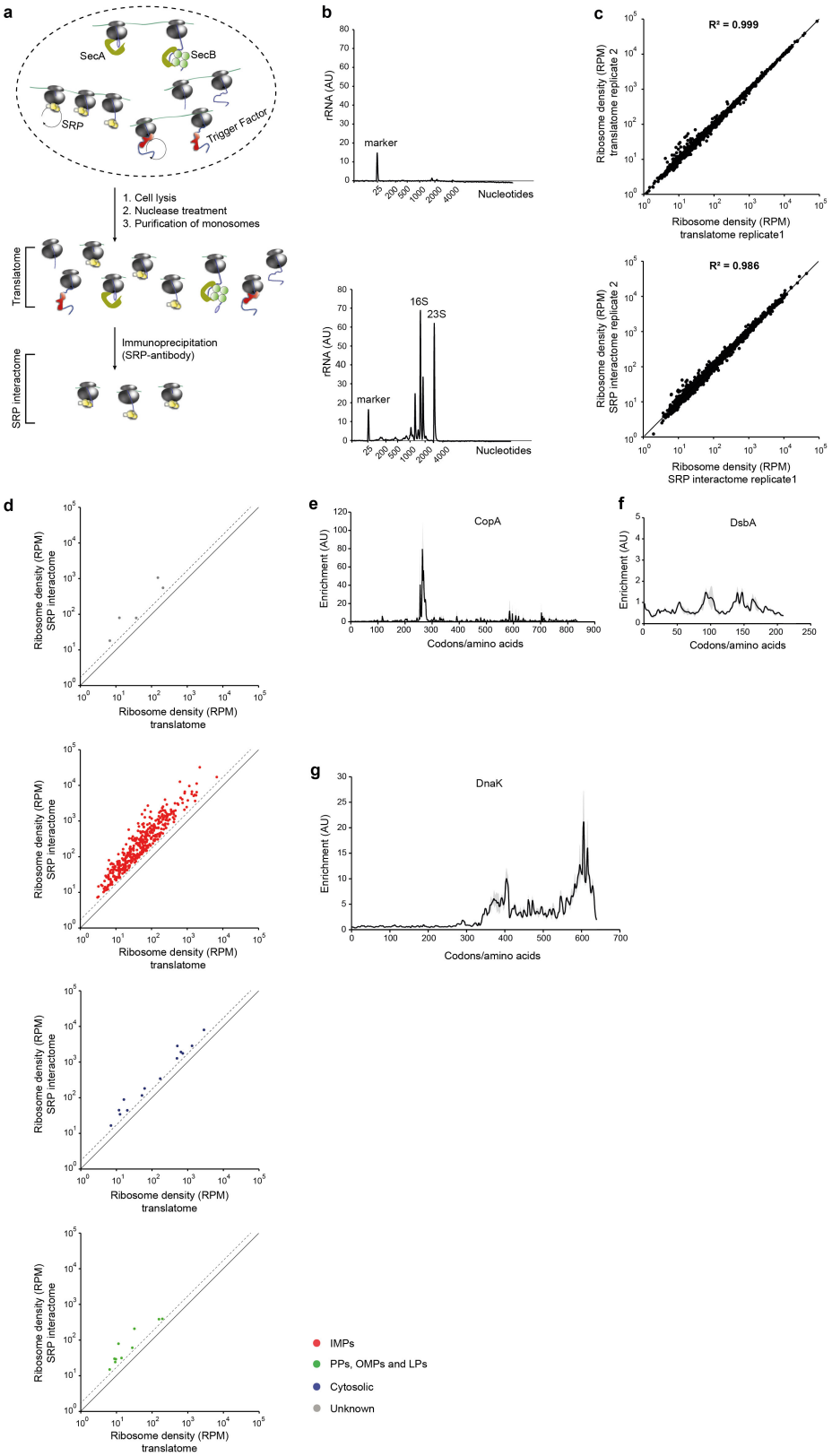
Peptide properties were computed by analysing amino acid content. Types of residues were collected from amino acid counts, combining D and E as acidic, I, L, M and V as aliphatic, F, W and Y as aromatic, H, K and R as basic, G and P as helix breaking and C, N, Q, S and T as polar. Average Kyte–Doolittle values, ΔG values, hydrophobic moments, isoelectric points, volumes, α-helical-, β-strand- and turn propensities were collected as a sum over the reference values for all contributing residues divided by the total number of residues in the analysed sequence. The percentage of sequences containing residues contributing to a consecutive hydrophobic stretch was also determined. In addition, signal peptides, which have a length of 23 residues on average, were analysed in a similar fashion and percentages of values with respect to the sequence length were computed.

Cloning of MsbA swap plasmid. A silently mutated DNA fragment encoding the N-terminal part of MsbA (bp 1–303) with swapped TMD1 and TMD2 was synthesized, fused with the 5′-end of *eyfp* and cloned in pTrc99B by In-Fusion cloning (Clonetech Laboratories).

Purification of RNCs for *in vitro* binding studies. Plasmids encoding N-terminal fragments of DsbA and MsbA fused to the N terminus of the minimal SecM stalling sequence were constructed using standard protocols and SecM encoding plasmids described previously[35]. Stalled RNCs contained N-terminal Strep-tagged Sumo fusion proteins. RNCs were generated in cells lacking trigger factor and purified by Strep-tag affinity purification followed by removal of the N-terminal Strep-Sumo tag using the protease fragment of the Sumo protease Ulp1 as described[35]. The nascent chain sequences remaining after purification are (signal sequence (SS) and TMD in bold): MsbA(**TMD1**)–SecM, MHNDKDLSTWQTFRRLWP**TIAPFKAGLIVAGVALILNAA**SDTFMLSLLKPLLFSTPVWISQAQGIRAGP; MsbA(**TMD2**)–SecM, MDGFGKTD**RSVLVWMPLVVIGLMILRGITSY**VSSYCISWVSFSTPVWISQAQGIRAGP; DsbA(**SS**)–SecM, M**KKIWLALAGLV LAFSASAAQ**YEDGKQYTTLEFSTPVWISQAQGIRAGP.

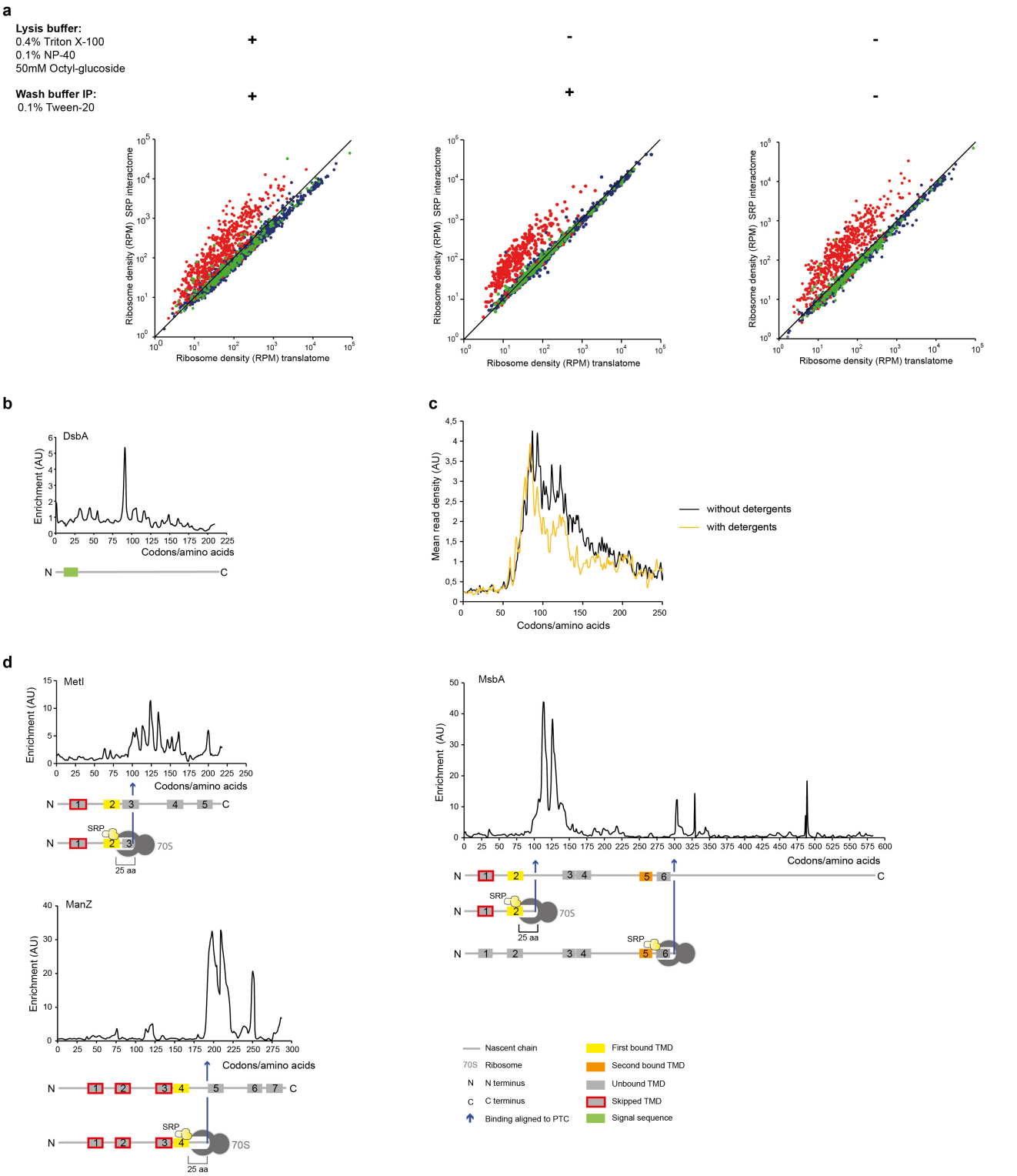
RNC *in vitro* binding studies. Purified SRP (1.5 μM) and RNCs (0.5 μM) were mixed and incubated for 15 min in low salt buffer (50 mM HEPES-KOH pH 7.0, 100 mM KOAc, 12 mM MgOAc, 1 mM DTT and Roche complete protease inhibitor). The salt concentration was adjusted and samples were loaded on 30% sucrose cushions prepared with salt-adjusted binding buffers. RNCs were sedimented by centrifugation for 75 min, 75,000 r.p.m., at 4 °C (Beckmann AT2 S100 rotor). Sixty microlitres of the supernatant was mixed with SDS sample buffer, the remaining supernatant was discarded and the pelleted ribosomes were resuspended. Samples were analysed by SDS–PAGE (4–20% gradient gels) followed by staining using SYPRO Ruby Protein Gel Stain (Thermo Fisher Scientific).

31. Casadaban, M. J. Transposition and fusion of the *lac* genes to selected promoters in *Escherichia coli* using bacteriophage lambda and Mu. *J. Mol. Biol.* **104,** 541–555 (1976).
32. Kramer, G. *et al.* Functional dissection of *Escherichia coli* trigger factor: unraveling the function of individual domains. *J. Bacteriol.* **186,** 3777–3784 (2004).
33. Becker, A. H., Oh, E., Weissman, J. S., Kramer, G. & Bukau, B. Selective ribosome profiling as a tool for studying the interaction of chaperones and targeting factors with nascent polypeptide chains and ribosomes. *Nature Protocols* **8,** 2212–2239 (2013).
34. Hessa, T. *et al.* Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* **433,** 377–381 (2005).
35. Rutkowska, A. *et al.* Large-scale purification of ribosome-nascent chain complexes for biochemical and structural studies. *FEBS Lett.* **583,** 2407–2413 (2009).

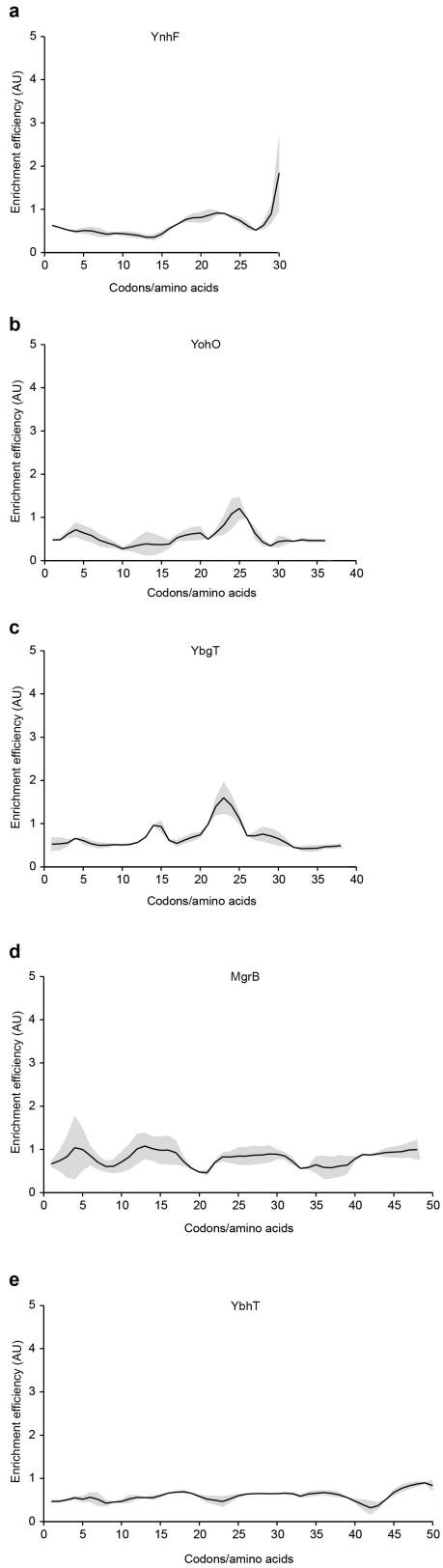


Extended Data Figure 1 | See next page for caption.

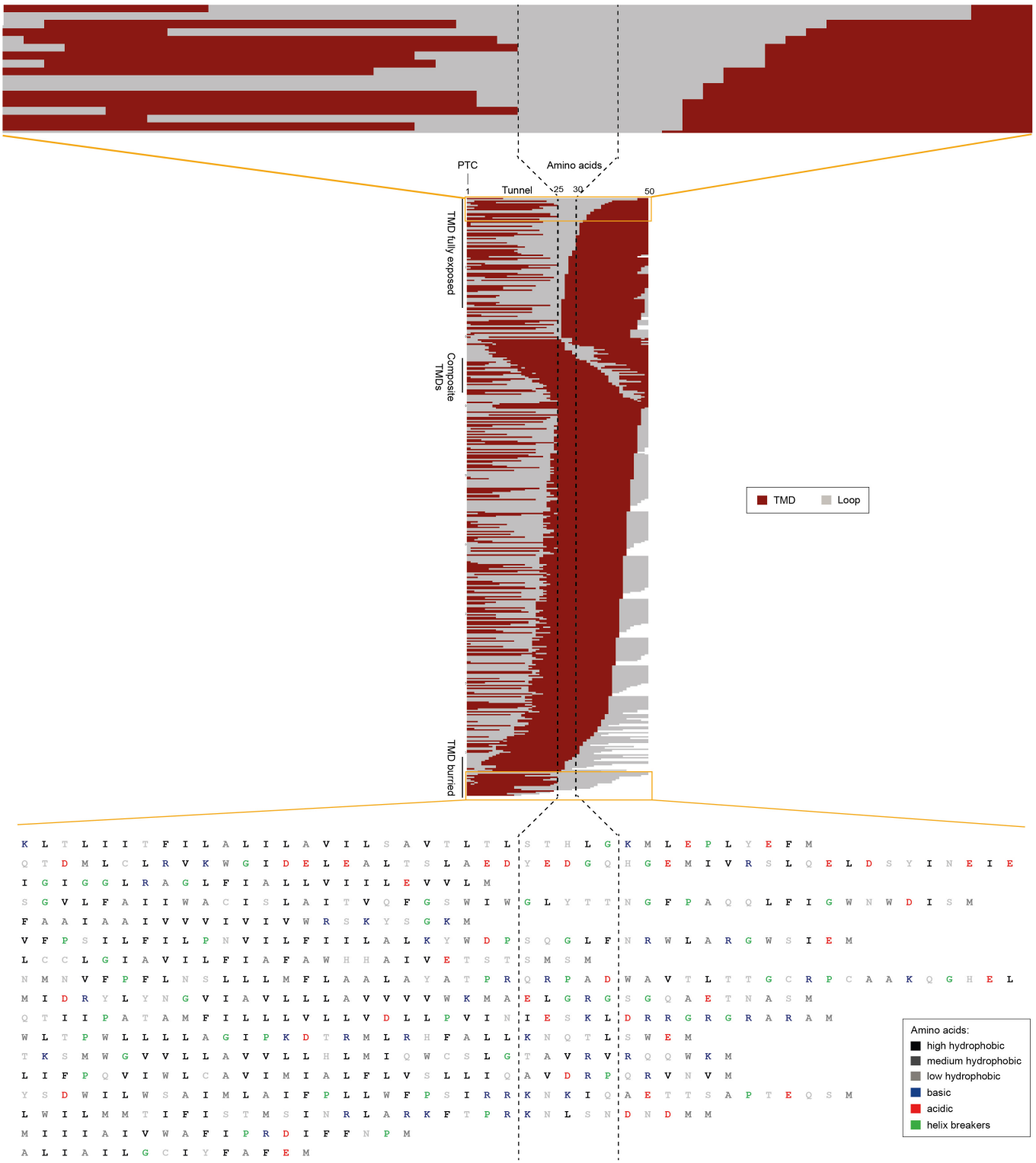
Extended Data Figure 1 | Selective ribosome profiling of *E. coli* SRP. **a**, Experimental scheme of selective ribosome profiling (SeRP) of *E. coli* SRP-bound RNCs. Cells were harvested in mid-log phase via rapid filtration, frozen in liquid nitrogen and lysed in a frozen state with a cryo mill. After thawing, polysomes were digested with micrococcal nuclease. Monosomes were purified by sucrose cushion centrifugation (translatome). SRP-bound RNCs were immunopurified using an SRP-specific polyclonal rabbit antibody (SRP interactome). **b**, Bioanalyzer spectra quantifying the amount of co-purified ribosomes in control immunoprecipitation (top) and SRP immunoprecipitation (bottom). The 16S ribosomal RNA (rRNA) of the small ribosomal subunit and the 23S rRNA of the large subunit are indicated. **c**, Reproducibility of translatome (left) and SRP interactome (right) data sets from biological replicates **d**, Gene expression levels of translatome and SRP interactome are compared. Only SRP substrates that pass a threshold of twofold enrichment are coloured according to localization (cytoplasm in blue, inner membrane in red, outer membrane, lipoproteins and periplasm in green, no localization known in grey). **e**, CopA ratio-enrichment profile of SRP interactome and translatome, Pearson correlation coefficient 0.74. Light grey shadows indicate the variation between two biological replicates. **f**, DsbA ratio-enrichment profile, Pearson correlation coefficient 0.67. Shadows as in **e**. **g**, DnaK ratio-enrichment profile of SRP interactome and translatome. Shadows as in **e**.



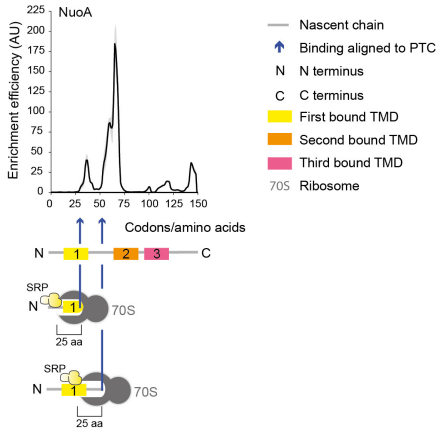
Extended Data Figure 2 | Selective ribosome profiling of *E. coli* SRP omitting detergents. a, Gene expression levels of the translatome and SRP interactome are compared for different experimental setups (with detergents in lysis and wash buffer ($n = 2$), detergent only in wash buffer ($n = 2$) and omitting detergents at all ($n = 1$)). ORFs are coloured according to localization (cytoplasm in blue, inner membrane in red, outer membrane, lipoproteins and periplasm in green, no localization known in grey). **b**, DsbA ratio-enrichment profile in the absence of detergents ($n = 1$). **c**, Metagene SRP interaction profile aligned to the N terminus of the initial TMD that is skipped in the presence of detergents (orange) and in the absence of detergents (black). **d**, Ratio-enrichment profiles in the absence of detergents of MetI, MsbA and ManZ ($n = 1$).



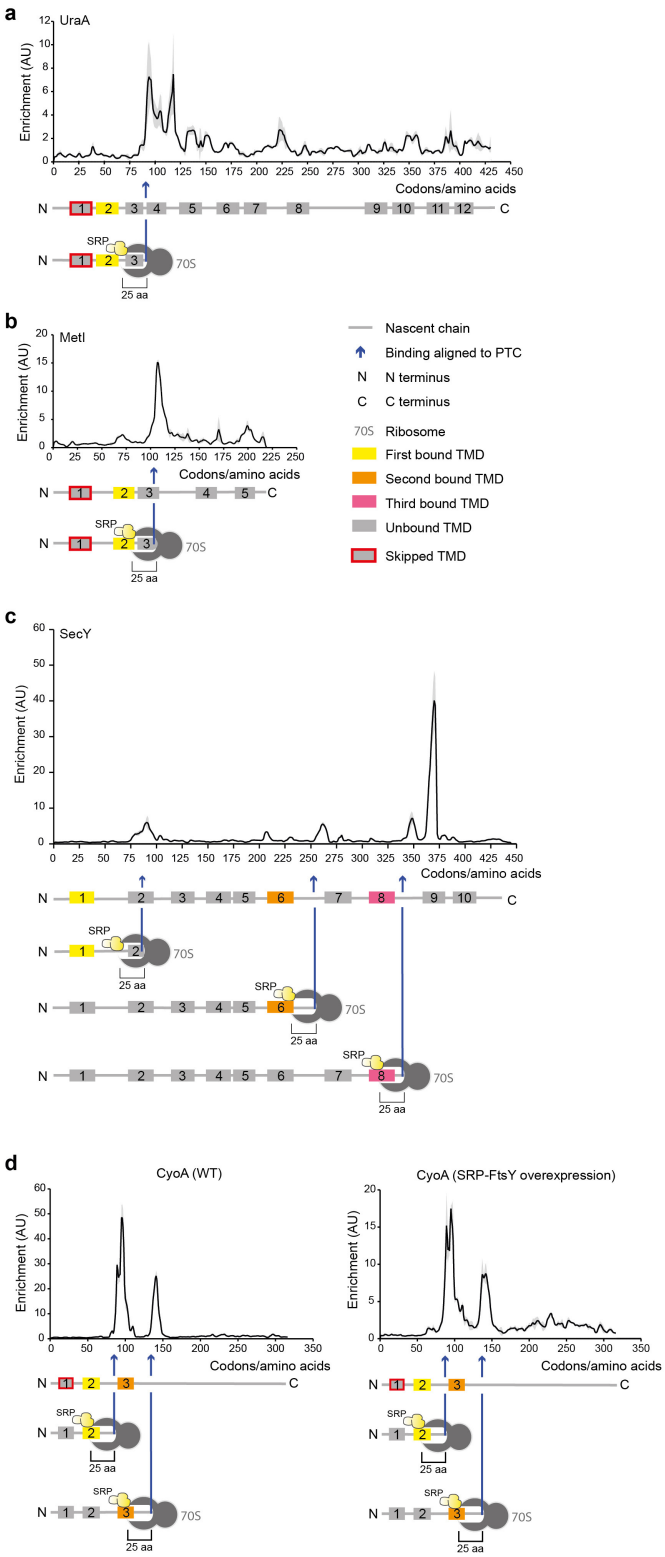
Extended Data Figure 3 | Interaction profiles of SRP with nascent YnhF, YohO, YbgT, MgrB and YbhT. Light grey shading indicates the variation between two biological replicates.



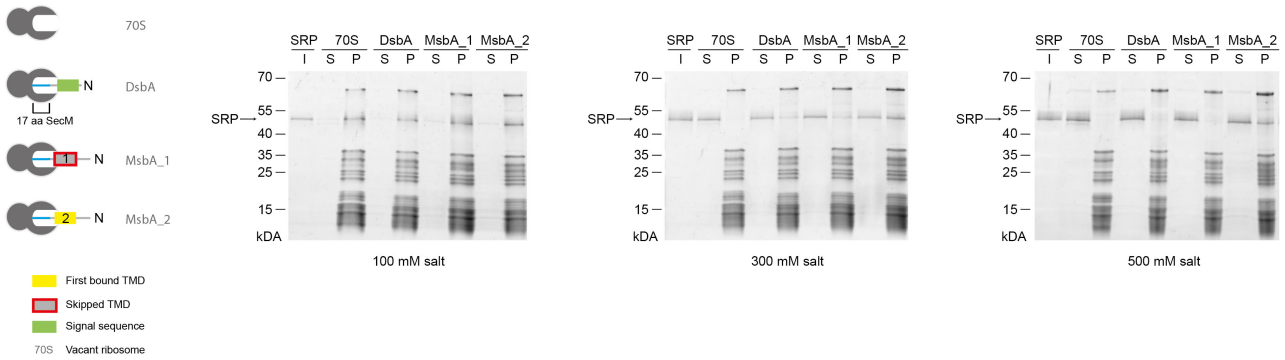
Extended Data Figure 4 | Heatmap representation of TMD positioning of inner membrane SRP substrates at the time point of first SRP binding. TMDs are shown in dark red and segments (loops) located outside the membrane bilayer are shown in light grey, dashed lines indicate the area of the ribosomal tunnel exit. Substrates that are bound by SRP without exposing a TMD near the ribosome surface are shown at amino acid resolution. Amino acid colour code: highly hydrophobic amino acids, black; medium hydrophobic amino acids, dark grey; low hydrophobic amino acids, light grey; basic amino acids, blue; acidic amino acids, red; helix breakers, green.



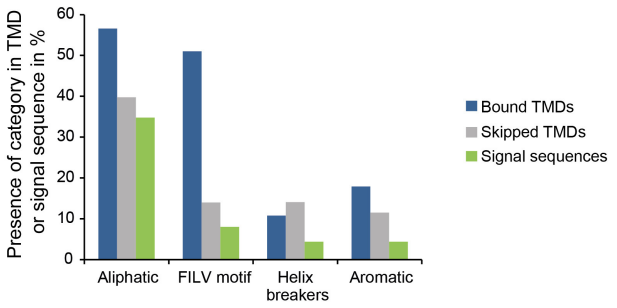
Extended Data Figure 5 | NuoA ratio enrichment profile of SRP binding. Binding events are correlated with topology. SRP binds nascent NuoA when the first TMD is still buried in the tunnel. Light grey shadows indicate the variation between two biological replicates.



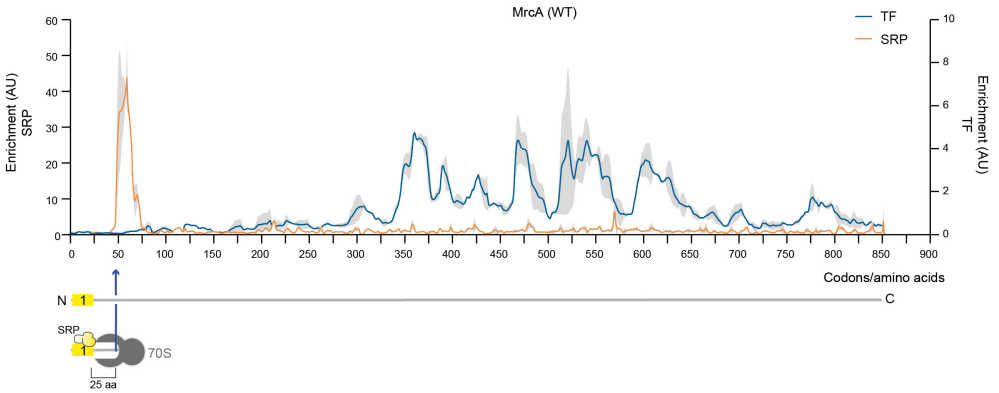
Extended Data Figure 6 | SRP interaction profiles. a–c, SRP interaction profile with nascent UraA (**a**), MetI (**b**) and SecY (**c**). SRP-binding peaks are correlated with protein topology. Light grey shading indicates variation between two biological replicates. **d**, SRP interaction profile with nascent CyoA in wild-type (WT) cells and cells overexpressing SRP and FtsY. Shading as in **a–c**.



Extended Data Figure 7 | SRP interaction with purified RNCs at different salt concentration. Left, cartoons illustrating the composition of nascent chains. Right, SRP–RNC interaction at indicated salt (ammonium acetate) concentrations analysed by sucrose cushion centrifugation and SDS–PAGE. Proteins are visualized using SYPRO Ruby Protein Gel Stain (Thermo Fisher Scientific). I, purified SRP; P, ribosomal pellet fraction; S, supernatant.



Extended Data Figure 8 | Quantification of ribosome-exposed nascent-chain binding properties of SRP. Four categories of binding markers (aliphatic amino acids, FILV motif, helix breakers and aromatic amino acids) quantified in bound TMDs, skipped TMDs and signal sequences.



Extended Data Figure 9 | Interaction profile of TF and SRP with nascent MrcA. TF, blue; SRP, orange. Light grey shading indicates the variation between two biological replicates.

LETTER

doi:10.1038/nature19309

Cotranslational signal–independent SRP preloading during membrane targeting

Justin W. Chartron[1], Katherine C. L. Hunt[1] & Judith Frydman[1,2]

Ribosome-associated factors must properly decode the limited information available in nascent polypeptides to direct them to their correct cellular fate[1]. It is unclear how the low complexity information exposed by the nascent chain suffices for accurate recognition by the many factors competing for the limited surface near the ribosomal exit site[2,3]. Questions remain even for the well-studied cotranslational targeting cycle to the endoplasmic reticulum, involving recognition of linear hydrophobic signal sequences or transmembrane domains by the signal recognition particle (SRP)[4,5]. Notably, the SRP has low abundance relative to the large number of ribosome–nascent-chain complexes (RNCs), yet it accurately selects those destined for the endoplasmic reticulum[6]. Despite their overlapping specificities, the SRP and the cotranslationally acting Hsp70 display precise mutually exclusive selectivity *in vivo* for their cognate RNCs[7,8]. To understand cotranslational nascent chain recognition *in vivo*, here we investigate the cotranslational membrane-targeting cycle using ribosome profiling[9] in yeast cells coupled with biochemical fractionation of ribosome populations. We show that the SRP preferentially binds secretory RNCs before their targeting signals are translated. Non-coding mRNA elements can promote this signal-independent pre-recruitment of SRP. Our study defines the complex kinetic interaction between elongation in the cytosol and determinants in the polypeptide and mRNA that modulate SRP–substrate selection and membrane targeting.

Secretory proteins are proposed to target to the endoplasmic reticulum (ER) membrane either co- or post-translationally for subsequent translocation[10–12]. Mechanistic models of ER targeting and the role of the SRP derive primarily from cell-free systems using model proteins[10,13], raising the question of how these pathways function in the cell. To investigate membrane targeting *in vivo*, we fractionated soluble and membrane-attached ribosomes from yeast cells, and then used ribosome profiling (termed Ribo-seq)[9] to compare the ribosome-protected mRNA footprints from polysomes obtained from both fractions (Extended Data Fig. 1a). We derived a cotranslational membrane enrichment score for each coding sequence (Methods, Extended Data Fig. 1b and Supplementary Table 1). Transcripts encoding cytosolic or nuclear (cytonuclear) proteins were preferentially translated on cytosolic ribosomes and not enriched on membrane polysomes (Fig. 1a). Tail-anchored proteins, whose single or transmembrane domain (TMD) at the carboxyl terminus is only revealed posttranslationally[14], were also translated on cytosolic ribosomes. By contrast, many nuclear-encoded mitochondrial protein transcripts were enriched in the membrane-bound ribosome fraction, as expected[15]. Transcripts encoding ER-destined secretory proteins were highly enriched on membrane-bound ribosomes. Proteins containing a signal sequence (SS) or TMD had comparable cotranslational membrane enrichment, conflicting with the idea that the targeting signal itself distinguishes which proteins are targeted co- or post-translationally to the ER[11,12] (Fig. 1a).

Ribosome profiling provides a snapshot of the abundance of ribosomes at each codon of each mRNA[9], revealing the dynamics of translation on soluble versus membrane-bound ribosomes. For cytonuclear proteins, soluble ribosome-protected reads were distributed across the entire reading frame, consistent with complete translation in the cytosol (Extended Data Fig. 1c). For secretory proteins, both soluble and membrane-bound polysomes produced protected reads. Cytosolic translation represented only a small fraction of any given secretory transcript, and most of the secretory mRNA pool was membrane anchored. In the classical understanding of cotranslational targeting, secretory protein RNCs bind to the membrane only after exposing a targeting signal[4]. Thus, there should be fewer RNCs found on the membrane translating the portion of transcripts not yet

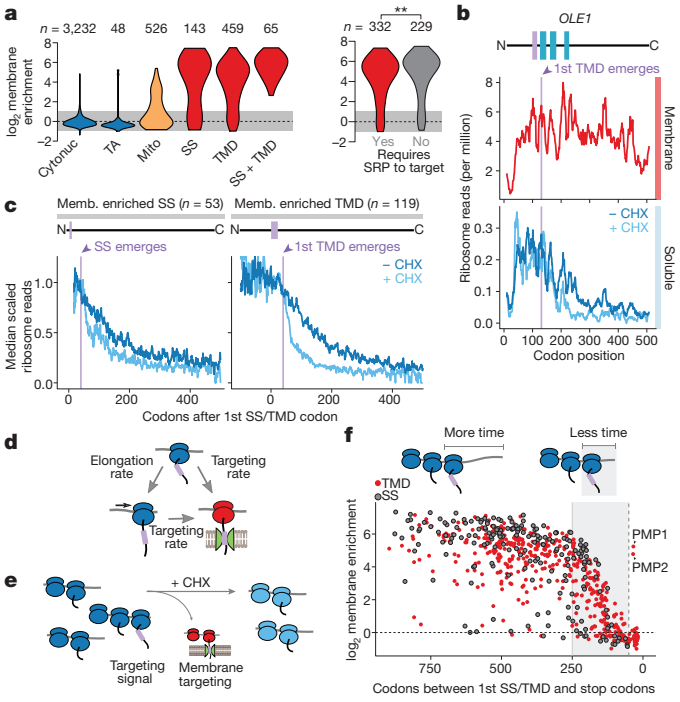


Figure 1 | Cotranslational membrane enrichment. a, Distributions of the open reading frame (ORF) enrichment of ribosome-protected reads in the membrane fraction compared to the soluble fraction. ORFs were alternatively classified by expected SRP dependence[11]. Values are the mean from two biological replicates. $^{**}P \leq 0.01$, Wilcoxon rank-sum test. **b**, Ribosome-protected reads at each codon of an example transmembrane protein *OLE1*. Membrane topology is indicated above, with the first TMD in lavender. **c**, Metagene analysis of soluble fraction polysome-protected reads from transcripts that were at least twofold membrane enriched. ORFs were aligned at the targeting signal and scaled. **d**, Cotranslational membrane targeting is in competition with elongation. **e**, Elongation inhibitors provide additional time for polysomes exposing a targeting signal to localize to the membrane. **f**, Membrane enrichment was limited by the length of the reading frame remaining after the encoding of targeting signals. The vertical dashed line indicates 50 codons.

[1]Department of Biology, Stanford University Stanford, California 94305, USA. [2]Department of Genetics, Stanford University Stanford, California 94305, USA.

targeted, that is, at codon positions upstream of the first SS or TMD. However, the membrane-bound ribosome-protected reads were evenly distributed across the entire transcript (Fig. 1b and Extended Data Fig. 1c, d). This suggests that once targeted, secretory mRNAs remain associated to the ER and their translation initiates at the membrane. This is consistent with the observed proximity of secretory RNCs to the translocon before synthesis of the targeting signal[16]. The small fraction of secretory mRNA in cytoplasmic pre-targeted RNCs probably represents the pioneer round of targeting.

The positioning of soluble ribosomes along mRNA provides insight into how secretory transcripts are targeted to the membrane. The highest read density for these messages mapped 5′ of the region encoding the first SS or TMD; read density declined after the first targeting signal was exposed by the ribosome, as expected from cotranslational signal-dependent targeting of soluble RNCs to the membrane (Fig. 1b, c and Extended Data Fig. 1d). Surprisingly, the loss of reads after signal emergence was gradual, resulting in many RNCs that remained soluble for hundreds of residues after SS or TMD exposure. This result was inconsistent with the elongation attenuation activity proposed for the SRP[17,18] and suggests that elongation continues on cytosolic RNCs upon exposure of a targeting signal (see Supplementary Discussion).

The idea that there is a kinetic competition between continuing elongation in the cytosol and RNC targeting to the membrane makes two testable predictions (Fig. 1d). First, pharmacological inhibition of elongation with cycloheximide (CHX) should decouple these processes, enhancing targeting of translocation-competent RNCs and promoting their depletion from the soluble fraction (Fig. 1e). Cells were subjected to a brief, two-minute CHX incubation before Ribo-seq analysis of soluble and membrane-bound polysomes. Importantly, such brief incubation did not perturb non-secretory polysomes (Extended Data Fig. 1c). By contrast, CHX treatment markedly reduced the soluble secretory reads, but only after cytosolic RNCs exposed the first SS or TMD, that is, 40 codons after its synthesis (Fig. 1b, c and Extended Data Fig. 1d).

The kinetic competition between targeting and elongation predicts that cotranslational membrane attachment is influenced by translation termination. In the absence of an elongation arrest, the probability of RNCs reaching the membrane cotranslationally will decrease as the first SS or TMD is found closer to the C terminus (Fig. 1f and Extended Data Fig. 1e). Indeed, we observed a decline in the maximum membrane enrichment of secretory RNCs when the first targeting signal is near the C terminus. Thus, secretory proteins with a late targeting signal, SS or TMD, must be targeted to the ER posttranslationally (Supplementary Discussion). Overall, our data suggest that cotranslational targeting to the ER in yeast is accomplished via a pioneer round of translation on soluble ribosomes that establishes a pool of ER-residing mRNA that initiate translation at the membrane (Extended Data Fig. 1f).

We next determined which RNCs are substrates of the SRP *in vivo*. Immunoprecipitation of Srp72p from total soluble RNCs was followed by ribosome profiling of both SRP-associated polysomes and monosomes (Fig. 2a and Extended Data Fig. 2a). Few transcripts encoding cytonuclear or mitochondrial proteins were enriched on SRP, confirming its specificity towards ER-destined transcripts. Notably, the SRP bound to all secretory RNCs that were cotranslationally targeted to the membrane, including SRP-dependent and SRP-independent proteins (Fig. 2b, c).

The number of ribosome-protected reads from soluble, SRP-bound transcripts diminished after ribosome exposure of the first SS or TMD, as expected from its targeting function (Fig. 2d). The loss was gradual and many SRP–RNCs remained soluble well after the targeting signal became fully exposed to the cytosol. This supports the notion that elongation proceeds on cytosolic ribosomes even after SRP binds, in contrast with the expected SRP-induced elongation arrest. Indeed, blocking elongation with CHX for 2 min before lysis caused a marked depletion in SRP-bound reads, but only for RNCs

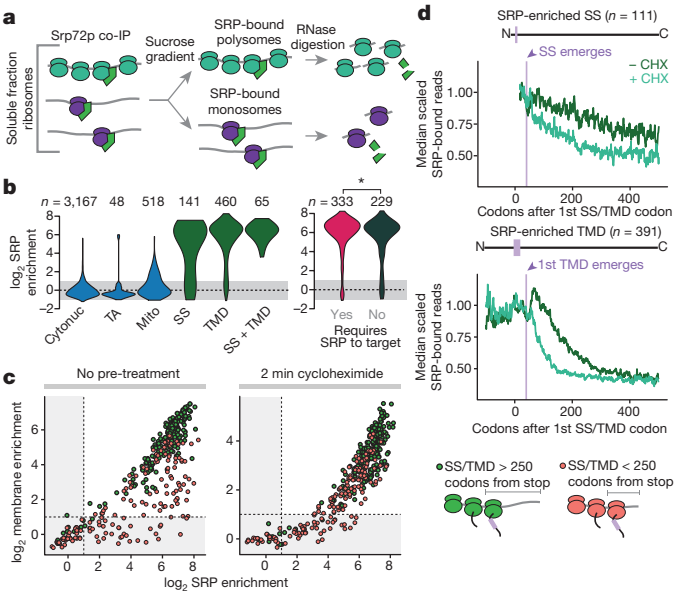


Figure 2 | Cotranslational enrichment of SRP. a, Srp72p-TAP was immunoprecipitated from the total soluble fraction. SRP-bound monosomes and polysomes were separated by sucrose gradient ultracentrifugation. **b**, Distributions of the ORF enrichment of ribosome-protected reads from SRP-bound soluble polysomes over the total soluble polysomes. ORFs were alternatively classified by expected SRP dependence[11]. Values are the mean from two biological replicates. TA, tail-anchored. $*P \leq 0.05$, Wilcoxon rank-sum test. **c**, Cotranslational membrane-fraction enrichment compared to SRP enrichment. **d**, Metagene analysis of soluble SRP-bound polysome-protected reads from transcripts that are at least twofold SRP-enriched. ORFs were aligned at the targeting signal and scaled.

exposing their first targeting signal (Fig. 2d). In principle, the delayed targeting of soluble RNCs to the membrane after SS/TMD emergence could reflect a delay in SRP binding rather than a lack of elongation arrest. Comparing the SRP and membrane enrichment to transcripts indicated that this is not the case. RNCs encoding late targeting signals, that is, near the C terminus, still bound SRP but did not target to the ER membrane (Supplementary Discussion, Fig. 2c and Extended Data Fig. 2b–d). Addition of CHX allowed these late-signal RNCs to enrich at the membrane, indicating the SRP–RNC complexes are competent for ER-targeting. We conclude that the SRP binds the nascent chain quickly, and continued elongation causes termination of late signals before targeting.

Although elongation arrest is not a general consequence of SRP binding *in vivo*, recent work showed that a rare-codon-directed slowdown of elongation facilitates SRP binding[19]. An intrinsic, non-SRP-dependent elongation slowdown should increase ribosome-protected reads at the same codon in both soluble SRP-bound and membrane-bound polysomes. Indeed, several transcripts presented such local increases in ribosome-protected reads at sites corresponding to exposure of a targeting signal on the ribosome (Extended Data Fig. 3a–c). Distinct elongation attenuation mechanisms observed at these sites included clusters of rare codons[19] and stalling polypeptide elements, such as stretches of positively charged amino acids, or proline motifs, positioned within the exit tunnel[20,21]. While most secretory transcripts were not significantly enriched in these attenuator elements compared to the proteome (Extended Data Fig. 3d, e), the few non-secretory proteins that cotranslationally bound to the SRP were enriched in elongation attenuation elements positioned at sites that exposed a near-cognate hydrophobic sequence for SRP binding (Extended Data Fig. 3d, f). We speculate that the presence of such elements enhances SRP recognition of the near-cognate hydrophobic tracts in these non-secretory proteins.

To understand the basis for the specificity of the SRP *in vivo*, we next determined the initial point of SRP recruitment to ribosomes

translating secretory proteins. Because polysomes require only a single SRP-bound ribosome to co-purify with Srp72p, additional strategies were necessary to identify mRNA footprints that originated from a single SRP-bound ribosome. We developed a protocol using *in vivo* monosomes to identify the initial SRP binding event on RNCs (Fig. 2a). At any given time, a fraction of transcripts contains only a single actively translating ribosome (Extended Data Fig. 4a). Total soluble monosomes yield a similar distribution of protected reads compared to polysomes (Extended Data Fig. 4b–e and Supplementary Discussion). We separated soluble SRP-bound monosomes from SRP-bound polysomes and subjected both fractions to Ribo-seq analysis (Extended Data Fig. 5a, b). Of note, the monosomes were necessarily bound to the SRP during the purification, and thus should reveal which codons are responsible for the initial SRP recruitment step.

The canonical model that the SRP recognizes the nascent chain after the targeting signal exits the ribosome[4] (Fig. 3a) makes several predictions. First, there should be few monosome-protected reads relative to polysomes before the first SS/TMD emerging from the ribosome tunnel; second, ribosome footprints should increase beginning approximately 40 codons after the first codon in the targeting signal, and third, monosome reads should decrease after full exposure of the SS/TMD, as SRP–RNCs are delivered to the membrane. Indeed, these patterns were observed in a subset of secretory transcripts with significantly more hydrophobic signals (Fig. 3b, Extended Data Figs 2e, f and 5c). SRP recruitment to these RNCs only occurred when the translated signals were fully exposed, and not while still in the exit tunnel[22,23] (Extended Data Fig. 5d and Supplementary Discussion).

Notably, most secretory transcripts did not conform to the predictions of the model (Fig. 3b, c and Extended Data Fig. 5e). Instead, ribosome footprints from most SRP-bound monosomes were abundant well before translation of the first targeting signal. For instance, the RNCs of *DAP2* were enriched on SRP from the start codon. For membrane proteins, SRP enrichment could be observed up to hundreds of codons before translation of the first TMD (Fig. 3d and Extended Data Fig. 5e). Thus, the exquisite selectivity of SRP towards secretory transcripts occurs via RNCs that have not yet translated any SS or TMD. Of note, the SRP-bound monosome reads did diminish upon full signal exposure by the ribosome. Thus, SRP is pre-recruited to secretory RNCs before the synthesis of an SS or TMD, but only after the emergence and recognition of the targeting signal can it promote membrane targeting, presumably owing to a conformational change in the SRP–RNC complex[22,24]. Our findings show that the SRP stably and preferentially binds ribosomes translating secretory mRNAs in a manner independent from the sequence of the exposed nascent chain (Fig. 3e). Models in which SRP scans all ribosomes with high affinity and rapid kinetics[25] do not explain our findings, as discussed in the Supplementary Discussion.

To begin to understand the determinants that confer specific recruitment of SRP without an exposed SS or TMD, we examined the most extreme cases of nascent-chain-independent SRP recruitment. *PMP1* and *PMP2* encode two abundant, small membrane proteins of 40 and 43 amino acids, respectively. Even though the entire proteins are smaller than the length of the ribosomal tunnel, *PMP1* and *PMP2* RNCs bind to the SRP throughout translation (Fig. 3f and Extended Data Fig. 6a, b).

We considered whether non-coding mRNA determinants could confer nascent-chain-independent SRP recruitment. *PMP1* and *PMP2* contain long 3′ untranslated regions (UTRs) implicated in membrane

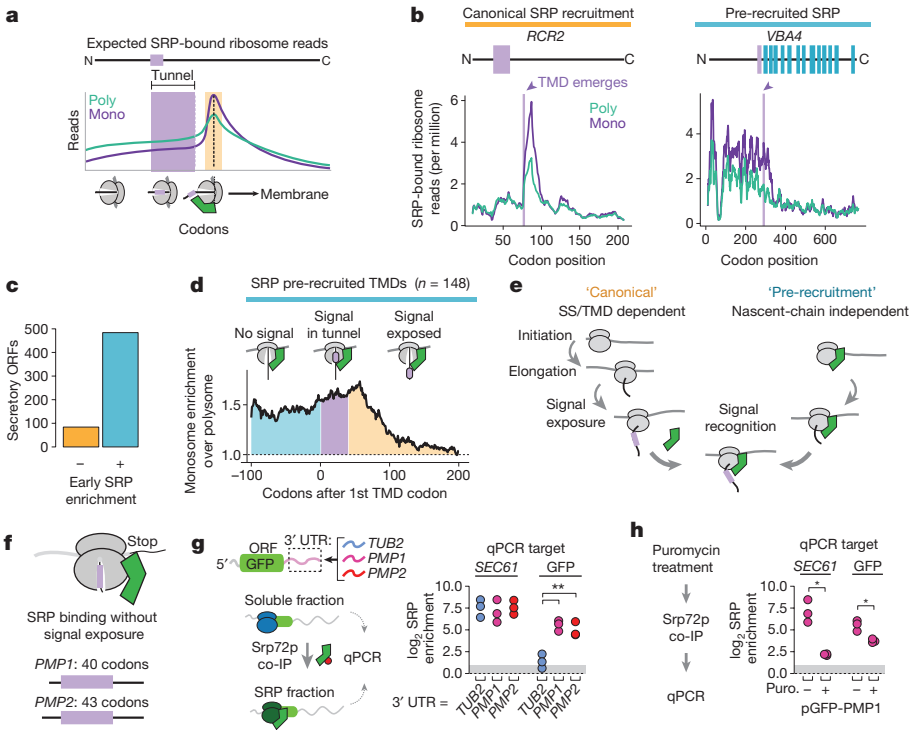


Figure 3 | Distinct mechanisms of SRP recruitment. a, Recruitment of SRP to RNCs is expected to increase ribosome-protected reads from SRP-bound monosomes when an SS or TMD is exposed to the cytosol (orange). **b**, Distributions of SRP-bound ribosome reads on representative transcripts from CHX-treated cultures. Selected transcripts are *RCR2* and *VBA4*. **c**, Most secretory proteins demonstrated SRP enrichment before signal exposure. **d**, Metagene plot of the median value of enrichment of SRP-bound monosomes compared to polysomes. Included transcripts encode TMDs at least 40 codons from the start codon. Shaded areas represent enrichment before the TMD is encoded (cyan), while the TMD is in the ribosome exit tunnel (lavender), and after the TMD is exposed (orange). **e**, Two mechanisms for SRP to select secretory mRNA. **f**, *PMP1* and *PMP2* were the only tail-anchored proteins that enriched SRP. **g**, The GFP ORF was fused to the indicated 3′ UTRs and expressed *in vivo*. Srp72p-TAP was immunoprecipitated from the total soluble fraction and RNAs were subject to quantitative PCR (qPCR). $n = 3$ biological replicates; $^{**}P \leq 0.01$, Welch's *t*-test. **h**, Puromycin treatment of lysate from yeast expressing GFP with the *PMP1* 3′ UTR was followed by SRP immunoprecipitation and qPCR. $n = 3$ biological replicates; $^{*}P \leq 0.05$, Welch's *t*-test.

attachment[26]. We thus tested the effect of fusing the 3′ UTR of either *PMP1* or *PMP2* to the mRNA of cytosolic green fluorescent protein (GFP) lacking any targeting signal (Fig. 3g). The 3′ UTR of cytosolic protein *TUB2* served as a control. Notably, the 3′ UTRs of either *PMP1* or *PMP2* conferred cotranslational SRP binding to the GFP transcripts, as well as membrane localization, whereas the 3′ UTR of *TUB2* did not (Fig. 3g, Extended Data Fig. 6c). For all constructs, GFP protein was diffuse and cytosolic, indicating that the 3′ UTR alone is insufficient to promote substantial translocation of GFP into the ER (Extended Data Fig. 6d). Notably, the 3′ UTR of endogenous *PMP1* is functionally important *in vivo*. Thus, replacing the 3′ UTR of the *PMP1* gene with the 3′ UTR of *TUB2* resulted in a growth defect more severe than complete deletion of the entire *PMP1* gene (Extended Data Fig. 6e). Perhaps mislocalization of the TMD in the absence of the 3′ UTR is more toxic than the loss of gene function.

Two non-exclusive models can account for SRP recruitment by the *PMP1* and *PMP2* 3′ UTRs. First, SRP binds to the mRNA, either directly or through other RNA-binding proteins. Alternatively, ribosomes translating *PMP1* or *PMP2* recruit SRP in a 3′ UTR-mediated manner (Extended Data Fig. 6f). To distinguish between these possibilities, a puromycin incubation was used to disrupt elongating[27] ribosomes before fractionation and SRP immunoprecipitation. This treatment caused a significant reduction in the GFP-PMP1 mRNA that copurified with SRP (Fig. 3h). Thus, translating ribosomes promote SRP recruitment to the GFP–PMP1 transcript. Of note, puromycin also disrupted the SRP interaction with the *SEC61* mRNA control. We thus next examined the general role of translation in SRP recruitment.

We assessed the global ribosome dependency of SRP binding to secretory transcripts using either puromycin or CHX incubations to disrupt or stabilize elongating ribosomes, respectively. Srp72p-bound transcripts isolated from the soluble fraction were examined by RNA-seq (Fig. 4a). SRP association with all secretory mRNAs was sensitive to puromycin. Transcripts that only recruit SRP through a canonical nascent chain interaction were more dependent on elongating ribosomes. The reduced puromycin sensitivity observed for pre-enriched transcripts may arise from the inability of puromycin to disrupt initiating ribosomes[27], which appear able to recruit SRP (Extended Data Fig. 6g).

We next examined whether the membrane association of secretory transcripts similarly depends on continuing translation. In principle, ER-localized proteins could recruit secretory transcripts to the membrane in the absence of translation[26,28] (Fig. 4b). Membrane and soluble mRNAs were fractionated in the presence and absence of puromycin treatment and subjected to RNA sequencing (RNA-seq) analysis (Extended Data Fig. 7). Disruption of translating ribosomes reduced membrane enrichment for all secretory protein transcripts, including those of *PMP1* and *PMP2*. This result was confirmed using the GFP-PMP1 reporter (Extended Data Fig. 6h).

The translation-dependence of membrane association for secretory transcripts was further examined using a temperature-sensitive allele of eIF3 subunit *PRT1*, *prt1-1*. Shifting cells to the non-permissive temperature precludes mRNA binding to the 40S subunit[29], allowing elongating ribosomes to run off (Fig. 4c). After displacing ribosomes from the mRNA, soluble and membrane fractions were analysed by RNA-seq. Notably, the only mRNAs that remained in the membrane fraction corresponded to mitochondrially encoded proteins. The membrane enrichment of all ER-destined secretory transcripts was abolished in the absence of translation (Fig. 4d). Thus, translation is required for the observed association of mRNAs with membranes.

Our findings define the principles of cotranslational membrane targeting and the role of SRP in this process and provide a solution to the paradox of how SRP achieves exquisite specificity *in vivo* despite its low abundance, its substrate binding promiscuity, and despite the competition from abundant cytosolic chaperones[30] that could potentially bind SSs or TMDs. For most mRNAs, SRP does not need to scan translating ribosomes rapidly for binding of

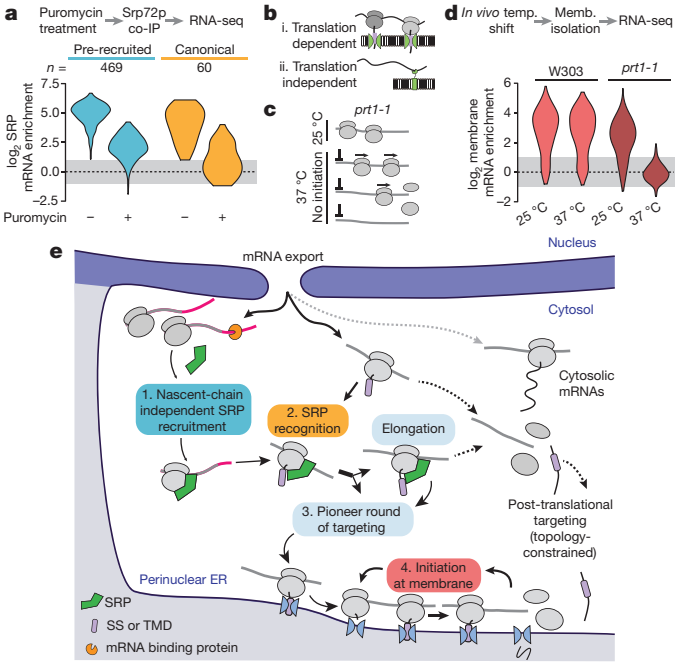


Figure 4 | Translation and the role of SRP. a, Distributions of RNA-seq SRP enrichment scores from secretory protein transcripts (SS, TMD, SS-TMD or tail-anchored), with or without puromycin treatment. Included ORFs have at least twofold SRP enrichment without puromycin. **b**, Transcripts are retained on the membrane through binding of the RNC to the translocon. It is also possible that mRNA binding proteins at the ER bind transcripts. **c**, The *prt1-1* allele prevents initiation at non-permissive temperatures. Translational run-off removes all ribosomes from transcripts. **d**, Distributions of RNA-seq membrane-enrichment scores of secretory protein transcripts ($n = 584$). **e**, After mRNA export, a pioneer round of targeting directs secretory transcripts to the ER membrane. SRP is specifically pre-recruited to transcripts that will present a functional targeting signal. After emergence of an SS or TMD, SRP directs RNCs to the ER membrane. Once at the ER membrane, transcripts are retained over several rounds of translation.

targeting sequences while ignoring near-cognate cytosolic hydrophobic sequences[6]. Instead, several mechanisms bias towards the correct SRP–RNC interactions (Fig. 4e). For most secretory mRNAs, SRP binds before targeting signals are synthesized, in a pioneer round of cytoplasmic translation. Pre-recruited SRP is thus poised to recognize the SS or TMD after emergence of a targeting signal from the ribosome[22] and facilitate membrane attachment. For a smaller fraction of clients with more hydrophobic-targeting signals, SRP recruitment is initiated by binding RNCs that fully expose SS or TMD in the nascent chain. We do not observe an SRP-induced elongation arrest, but some mRNAs have intrinsic elements attenuating elongation upon signal exposure. Since membrane targeting is in kinetic competition with continued elongation, posttranslational targeting dominates for proteins with a late targeting signal. Once at the membrane, secretory mRNAs remain bound through subsequent rounds of initiation and translocon engagement. These hydrophobic proteins will no longer compete with soluble proteins for cytosolic quality control components. Conversely, transcripts not captured in this first round of selection become more likely to encounter cytosolic chaperones. One important and surprising conclusion is that cotranslational events governing nascent polypeptide fate are not only guided by the nascent chain itself, but also rely on additional aspects of translation, such as mRNA itself and cellular organization. These findings illustrate the multi-layered nature of protein biogenesis fidelity.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 16 October 2015; accepted 18 July 2016.
Published online 3 August 2016.

1. Pechmann, S., Willmund, F. & Frydman, J. The ribosome as a hub for protein quality control. *Mol. Cell* **49,** 411–421 (2013).
2. Bornemann, T., Holtkamp, W. & Wintermeyer, W. Interplay between trigger factor and other protein biogenesis factors on the ribosome. *Nat. Commun.* **5,** 4180 (2014).
3. Nyathi, Y. & Pool, M. R. Analysis of the interplay of protein biogenesis factors at the ribosome exit site reveals new role for NAC. *J. Cell Biol.* **210,** 287–301 (2015).
4. Akopian, D., Shen, K., Zhang, X. & Shan, S. O. Signal recognition particle: an essential protein-targeting machine. *Annu. Rev. Biochem.* **82,** 693–721 (2013).
5. Zhang, X. & Shan, S. O. Fidelity of cotranslational protein targeting by the signal recognition particle. *Annu. Rev. Biophys.* **43,** 381–408 (2014).
6. Ogg, S. C. & Walter, P. SRP samples nascent chains for the presence of signal sequences by interacting with ribosomes at a discrete step during translation elongation. *Cell* **81,** 1075–1084 (1995).
7. Willmund, F. *et al.* The cotranslational function of ribosome-associated Hsp70 in eukaryotic protein homeostasis. *Cell* **152,** 196–209 (2013).
8. del Alamo, M. *et al.* Defining the specificity of cotranslationally acting chaperones by systematic analysis of mRNAs associated with ribosome-nascent chain complexes. *PLoS Biol.* **9,** e1001100 (2011).
9. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* **324,** 218–223 (2009).
10. Rapoport, T. A., Matlack, K. E., Plath, K., Misselwitz, B. & Staeck, O. Posttranslational protein translocation across the membrane of the endoplasmic reticulum. *Biol. Chem.* **380,** 1143–1150 (1999).
11. Ast, T., Cohen, G. & Schuldiner, M. A network of cytosolic factors targets SRP-independent proteins to the endoplasmic reticulum. *Cell* **152,** 1134–1145 (2013).
12. Ng, D. T., Brown, J. D. & Walter, P. Signal sequences specify the targeting route to the endoplasmic reticulum membrane. *J. Cell Biol.* **134,** 269–278 (1996).
13. Walter, P. & Johnson, A. E. Signal sequence recognition and protein targeting to the endoplasmic reticulum membrane. *Annu. Rev. Cell Biol.* **10,** 87–119 (1994).
14. Kutay, U., Ahnert-Hilger, G., Hartmann, E., Wiedenmann, B. & Rapoport, T. A. Transport route for synaptobrevin via a novel pathway of insertion into the endoplasmic reticulum membrane. *EMBO J.* **14,** 217–223 (1995).
15. Williams, C. C., Jan, C. H. & Weissman, J. S. Targeting and plasticity of mitochondrial proteins revealed by proximity-specific ribosome profiling. *Science* **346,** 748–751 (2014).
16. Jan, C. H., Williams, C. C. & Weissman, J. S. Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling. *Science* **346,** 1257521 (2014).
17. Lakkaraju, A. K., Mary, C., Scherrer, A., Johnson, A. E. & Strub, K. SRP keeps polypeptides translocation-competent by slowing translation to match limiting ER-targeting sites. *Cell* **133,** 440–451 (2008).
18. Mason, N., Ciufo, L. F. & Brown, J. D. Elongation arrest is a physiologically important function of signal recognition particle. *EMBO J.* **19,** 4164–4174 (2000).
19. Pechmann, S., Chartron, J. W. & Frydman, J. Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP *in vivo*. *Nat. Struct. Mol. Biol.* **21,** 1100–1105 (2014).
20. Lu, J. & Deutsch, C. Electrostatics in the ribosomal tunnel modulate chain elongation rates. *J. Mol. Biol.* **384,** 73–86 (2008).
21. Woolstenhulme, C. J., Guydosh, N. R., Green, R. & Buskirk, A. R. High-precision analysis of translational pausing by ribosome profiling in bacteria lacking EFP. *Cell Reports* **11,** 13–21 (2015).
22. Voorhees, R. M. & Hegde, R. S. Structures of the scanning and engaged states of the mammalian SRP-ribosome complex. *eLife* **4,** e07975 (2015).
23. Berndt, U., Oellerer, S., Zhang, Y., Johnson, A. E. & Rospert, S. A signal-anchor sequence stimulates signal recognition particle binding to ribosomes from inside the exit tunnel. *Proc. Natl Acad. Sci. USA* **106,** 1398–1403 (2009).
24. Hainzl, T. & Sauer-Eriksson, A. E. Signal-sequence induced conformational changes in the signal recognition particle. *Nat. Commun.* **6,** 7163 (2015).
25. Holtkamp, W. *et al.* Dynamic switch of the signal recognition particle from scanning to targeting. *Nat. Struct. Mol. Biol.* **19,** 1332–1337 (2012).
26. Loya, A. *et al.* The 3′-UTR mediates the cellular localization of an mRNA encoding a short plasma membrane protein. *RNA* **14,** 1352–1365 (2008).
27. Gao, X. *et al.* Quantitative profiling of initiating ribosomes *in vivo*. *Nat. Methods* **12,** 147–153 (2015).
28. Kraut-Cohen, J. *et al.* Translation- and SRP-independent mRNA targeting to the endoplasmic reticulum in the yeast *Saccharomyces cerevisiae*. *Mol. Biol. Cell* **24,** 3069–3084 (2013).
29. Naranda, T., MacMillan, S. E. & Hershey, J. W. Purified yeast translational initiation factor eIF-3 is an RNA-binding protein complex that contains the PRT1 protein. *J. Biol. Chem.* **269,** 32286–32292 (1994).
30. Elvekrog, M. M. & Walter, P. Dynamics of co-translational protein targeting. *Curr. Opin. Chem. Biol.* **29,** 79–86 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank P. Walter, J. S. Weissman and C. Jan for discussions; R. Andino and R. Hegde for critical reading of the manuscript. We thank S. Pechmann, K. M. Dalton, E. M. Sontag, P. T. Dolan and other members of the Frydman laboratory for advice on analysis. Sequencing was performed at the UCSF Center for Advanced Technology with assistance from E. Chow, J. Lund and A. Acevedo. J.W.C. is supported by an NIH NRSA award. This work was additionally supported by grants to J.F. from the NIH and HFSP.

Author Contributions J.W.C. and J.F. designed the study. K.C.L.H. performed experiments with *prt1-1*. J.W.C. performed all other experiments and analysis. J.W.C. and J.F. wrote the manuscript.

Author Information Data are deposited in Gene Expression Omnibus (GEO) under accession number GSE74393. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.F. (jfrydman@stanford.edu).

Reviewer Information *Nature* thanks R. Keenan and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment.

Yeast strains. Ribosome profiling (Ribo-seq) and qPCR assays were performed using BY4741 Srp72-TAP::His3MX, obtained from OpenBiosystems[31]. BY4741 SEC61-GFP::His3MX was obtained from Invitrogen[32]. Strain CY2522, containing the *prt1-1* temperature-sensitive allele[33], was provided by E. Craig. BY4741 PMP1Δ::kanMX4 was obtained from the yeast deletion library[34].

Ribosome profiling. For each biological replicate, six 500-ml cultures of YPD were grown with shaking at 30 °C to $OD_{600\,nm} = 0.8–1.0$, and collected one at a time by filtering through a 0.22-μm membrane. Cells were scraped off the filter in a single motion using a metal scoopula and then immersed in liquid nitrogen. When indicated, CHX treatments were performed by adding the drug to the culture to 100 μg ml^{-1} immediately before filtering. Filtration completed in approximately 2 min, and cells were scraped and immersed in liquid nitrogen within 3 s.

Lysis buffer comprised 50 mM potassium MOPS, pH 7.2, 275 mM potassium glutamate, 5 mM magnesium acetate, 1 mM DTT, 100 μg ml^{-1} CHX, and 20 U ml^{-1} Superase•In (Ambion). Two 3-ml aliquots were supplemented with Complete Protease Inhibitor Cocktail, EDTA-free (Roche) and frozen dropwise in liquid nitrogen. One 3-ml aliquot of frozen lysis buffer was combined with cells from 1.5 l of culture in a 50 ml ball mill chamber chilled in liquid nitrogen (Retsch). Cells were pulverized for 1 min at 20 Hz in a MM-301 mixer mill. Pulverized cells from 3 l of culture were combined and thawed in a room temperature water bath. Lysates were immediately centrifuged in a Type 70.1 Ti rotor (Beckman) for 10 min at 12,000 r.p.m. The following were then added to the supernatant: Triton X-100 to 0.01%, heparin sulfate to 0.2 mg ml^{-1}, and PMSF to 1 mM. Heparin was added as an RNase inhibitor only after fractionation as it may dislodge ribosomes from the membrane[35]. A low concentration of Triton X-100 reduces bead clumping during immunoprecipitation, and prevents aggregation upon elution. A portion of the supernatant was retained as the total soluble fraction. Three millilitres of lysis buffer supplemented with Triton X-100 to 1% and heparin sulfate to 0.2 mg ml^{-1} were added to the pellets. Pellets were resuspended using a glass dounce homogenizer fit to the internal diameter of the tube. Membrane extracts were centrifuged as before, and the detergent-extracted supernatant is recovered as the total membrane-bound fraction.

One millilitre of streptavidin-conjugated magnetic beads (Pierce) was saturated with biotinylated total rabbit IgG (Calbiochem). Beads were incubated with the total soluble fraction for 1 h at 4 °C. Beads were then washed three times with 1 ml of wash buffer, which consisted of lysis buffer supplemented with 0.2 mg ml^{-1} heparin and 0.01% Triton X-100. Beads were then incubated with 100 μl wash buffer, 2 μl Superase•In and 3 μl AcTEV protease (Invitrogen) for 1 h at room temperature. The first eluate was retained on ice, the digest was repeated, and the eluates combined, yielding the total SRP-bound fraction. The eluate was immediately quantified using $A_{260\,nm}$, and the total soluble and membrane-bound fractions were diluted in wash buffer to equivalent concentrations in 200 μl. Samples were layered on a 7–47% sucrose gradient prepared in wash buffer omitting RNase inhibitors. Gradients were centrifuged in a SW 41 Ti rotor (Beckman) for 2.5 h at 39,000 r.p.m., and fractionated using a UA-6 detector and Foxy Jr. fraction collector (ISCO). Five 1-ml fractions containing the polysomes were combined, as were two 1-ml fractions containing the monosomes. Samples were diluted to 6 ml in wash buffer without RNase inhibitors and centrifuged for 12 h at 50,000 r.p.m. in a Type 70.1 Ti rotor.

Ribosome pellets were resuspended in 250 μl of cutting buffer of 20 mM Tris, pH 7.5, 140 mM potassium chloride, 1.5 mM magnesium chloride, and 0.01% Triton X-100. Concentrations were typically 10–100 ng μl^{-1} as measured by $A_{260\,nm}$. Samples with greater concentration were diluted to 100 ng μl^{-1} in 250 μl. Fifty units of RNase I (Ambion) were added to each sample, and digests proceeded for 1 h at room temperature. Digests were stopped with the addition of 2 μl Superase•In, transferred to a MLA-130 centrifuge tube (Beckman) and underlaid with 750 μl of 35% sucrose in cutting buffer. Ribosomes were pelleted by centrifugation for 4.5 h at 70,000 r.p.m. Total RNA was extracted from the pellets using a miRNeasy kit (Qiagen). Libraries were prepared as previously described[36], quantified by qPCR (Kapa Biosciences), and sequenced using a HiSeq 2500 (Illumina).

RNA-seq. To test disruption of elongating ribosomes, yeast from one 500-ml culture of BY4741 Srp72-TAP::His3MX were lysed in ribosome profiling lysis buffer, which includes CHX. Yeast from a second culture were lysed in buffer prepared without CHX that was supplemented with 0.5 mM puromycin, 1 mM ATP, 1 mM GTP, 10 mM creatine phosphate and 40 μg ml^{-1} creatine kinase. Both lysates were incubated at room temperature for 10 min after thawing.

To test the effect of initiation *in vivo*, two 500-ml cultures of CY2522 (*prt1-1*) and two 500-ml cultures of W303 were grown with shaking at 25 °C to $OD_{600\,nm} = 0.6$. One culture of each strain was shifted to 37 °C, and the four cultures were grown one additional hour followed by harvesting by fast filtration.

Samples of the total soluble, total membrane, and total SRP-bound fractions were prepared as described for ribosome profiling. RNA was extracted using the hot SDS-phenol-chloroform method, and mRNA was purified using oligo-dT beads according to the manufacturer's instructions (NEB). Eluted mRNA was then fragmented under alkaline conditions[9], and fragments of 35–50 nucleotides were purified by PAGE. Libraries were constructed as for Ribo-seq and sequenced using a HiSeq 4000 (Illumina).

Data processing for enrichment scores. Adaptor sequences were trimmed from sequencing reads using Cutadapt[37]. Two rounds of alignment are performed using Bowtie, and Tophat[38,39]. First, sequences are aligned against a library comprising mature ribosomal rRNA with Bowtie. Unaligned reads are retained and then aligned against S288C Release 64-1-1 (http://www.yeastgenome.org) using Tophat. Any read with more than one match was removed, and reads were assigned to ORFs and counted using the GenomicAlignments package in Bioconductor[40,41]. Dubious ORFs were omitted.

Identification of targeting signals and classification of ORFs. Different SS and TMD prediction programs vary in their output, and so we included only consistently predicted targeting signals in our analysis. Protein sequences were given to SignalP V3 (ref. 42), Phobius[43], Philius[44], and TMHMM[45]. The following scheme was used for classification.

Mitochondrial. All mitochondrially encoded proteins, and nuclear-encoded proteins that localized to this organelle in at least one of two fluorescence-based screens[32,46].

SSs. SSs from non-mitochondrial proteins predicted by SignalP, Phobius and Philius that did not contain any TMDs predicted by Phobius or Philius. The first residue of the hydrophobic domain, as predicted by Phobius, was designated as the first signal residue. Predictions by TMHMM within the first 50 residues were ignored. SSs were defined as 'looped' if they enriched on Ssh1p at least 90 codons after the first SS codon[16]. Every GPI-anchored protein (as previously annotated[11]) satisfies conditions for an SS, and was included. Some GPI-anchored proteins have predicted TMDs near the stop codon; these do not expose cotranslationally.

TMDs. The first TMDs of non-mitochondrial proteins were predicted by TMHMM, Phobius, and Philius within five amino acids for each pair of predictors (that is, Philius versus Phobius, Philius versus TMHMM, Phobius versus TMHMM). The first signal location was designated as the average of the three predictions, rounded down. If the TMD was within the first 50 codons, predictions by SignalP are ignored. Phobius and Philius did not predict a cleavable SS.

Tail-anchored. First TMDs that begin within 50 amino acids of the stop codon were designated tail-anchored.

Signal sequences with transmembrane domains (SS-TMD). SSs were predicted as above with at least one TMD predicted by both Phobius and Philius.

Cytonuclear. ORFs that had no predicted SS or TMD by SignalP, TMHMM, Phobius, or Philius, and which did not appear in the mitochondria. Since only fluorescence localization data were used to designate mitochondrial proteins, this set includes some true mitochondrial proteins.

Exceptions. All remaining sequences had an SS or TMD predicted by SignalP, Phobius, Philius, or TMHMM that was not predicted by the other programs. Because of the ambiguity in type and location of the targeting signal, these proteins are excluded from our analysis. Other exceptions included proteins with predicted TMDs from position 50 or later, as well as a SignalP SS prediction but no Phobius or Philius SS. We considered this ambiguity in the prediction of an SS, and excluded these ORFs.

Enrichment analysis. Count-based sequencing assays report changes in a transcript's abundance as changes in its proportion of the total sequencing reads. Thus, when a subset of transcripts is enriched in a sample, the proportion of reads from all other transcripts decreases by a corresponding amount. Here we assume that enrichment on the membrane or SRP represents active selection of certain transcripts, and the depletion of all others is passive. In other words, we assume that few transcripts will be specifically prevented from appearing in the membrane or SRP. The distributions of the 'Cytonuclear' sets in Figs 1a and 2b, which skew towards enrichment, are consistent with this assumption. Thus, we expect the distribution of non-enriched transcripts to be similar to their overall expression or translation. This makes direct comparison of different enrichment scores (that is, SRP enrichment vs membrane enrichment, under various drug treatments etc.) calculated from proportional abundance[15,16,47] unintuitive, because a component of enrichment appears as depletion of non-enriched mRNA.

In current approaches for differential gene expression, most genes are assumed to have unaltered abundance, and library sizes are normalized by a robust estimator such as median ratio method[48] and trimmed mean of M-values (TMM)[49]. However, we expected changes for up to a third of ORFs. We used the TMM method of DESeq to derive library scale factors using reference ORFs, selected as

those designated as 'cytonuclear'. We applied these scale factors to the counts for every ORF, and calculated enrichment as the ratio of scaled reads between sets. Biological replicates were scaled separately and then averaged. The robust nature of scale estimation allowed for extreme cases of the reference set to have high enrichment scores. Scores were only reported for ORFs that have at least 100 total reads between replicates in each of the compared samples[9,47].

Figure 2c included proteins designated as SS, TMD and SS-TMD that have at least 100 reads in all four data sets, with the following exceptions: YEL050C, YLR077W, YML061C, YOL053W and YPL132W, which were all observed on the surface on the mitochondria[15] and had at least fourfold membrane enrichment here. Membrane enrichment after CHX treatment was determined using total ribosome-protected reads from the soluble and membrane fractions.

Mapping of ribosome-protected reads to codons. Reads were mapped to codons using an alternate method. After filtering rRNA, reads were aligned to a Bowtie library comprising coding sequences, plus the stop codon and 21-nucleotides flanking upstream and downstream. Using combined data from the SRP-pulldown and membrane polysome replicates, ORFs for which at least 20% of reads could map to a second ORF were removed, leaving a working set of 5,441 genes. Footprints of 26–35 nucleotides were processed separately for each length. The nucleotide that mapped to the centre of each read (rounded down) was given a value of 1, and reads were summed at each nucleotide position. A metagene analysis was performed[50] and for each footprint length, an integer offset was determined so that the characteristically large peak at the start codon was maximized at the second nucleotide position (that is, aTg). Then, reads of all lengths were offset and combined. Nucleotide reads were summed for each codon.

Read distributions. For each sample, the total reads from elongating ribosomes were determined by adding counts from all codons excluding the first two and last two sense codons. Reads from biological replicates were summed to increase overall read depth, but owing to the high overall reproducibility, all of our conclusions can be demonstrated by treating replicates separately. The reads at each codon position are then divided by this total and multiplied by one million to yield reads per million (RPM). Values at each codon are smoothed using an 11 residue rolling average. The positions of TMDs in topology diagrams were taken from the TMHMM prediction. We caution that predictors may differ in the number and position of subsequence TMDs. Positions of SSs are the H-region predicted by Phobius. The point at which an SS or TMD begins to emerge is considered 40 codons after the first encoded residue of the signal.

For metagene plots, reads at each codon are smoothed using a 5 residue rolling average. ORFs are then aligned as indicated, and the median and interquartile ranges are calculated at each position. For each ORF in Figs 1c and 2d, reads at each codon position are divided by the mean reads per codon within the range +20 to +40 after first signal codon. Included ORFs have at least 20 reads within this window in each data set shown. The first 30 codons of each ORF are excluded to avoid the universally observed low-density region near the start codon.

Identification of SRP recruitment to the nascent chain. Increases in ribosome-protected reads in the soluble SRP-bound polysome data set were observed for a subset of ORFs at codon positions coincident with the exposure of targeting signals. We developed a clustering scheme that sorted ORFs by the shape of the distribution of ribosome-protected reads specific to SRP-bound polysomes. The test set comprised 568 SS, TMD or SS-TMD proteins with at an average of at least 3 reads per codon in both soluble SRP-bound polysome and membrane-bound polysome sets. For each ORF, we first smooth read counts from each data set with a leading 15-residue moving average window. We corrected for local features intrinsic to the sequence (that is, appearing in all fractions) by dividing the smoothed SRP-bound polysome reads by the smoothed membrane-bound polysome reads at each codon; positions with fewer than 3 reads in either smoothed set were omitted. Peak codon positions were identified as the maximum value within 30–180 residues after the first predicted targeting signal codon. Each ORF was scaled by dividing the value at each codon position by the mean value over the range from 50 codons before to 200 codons after the peak. Codon positions outside this range are discarded. Scaled values are then used to generate an empirical cumulative distribution function (ECDF). The ECDF is sampled from 0.0 to 2.0 in 0.1 steps. The samplings from the ECDFs were used for agglomerative hierarchical clustering using a Euclidean distance function and Ward's minimum variance method[51]. The first split in the population distinguished ORFs having strong peaks from ORFs with weak or no peaks.

To analyse the distances between the first signal codon and the peak, peaks must be unambiguously assigned to the first targeting signal; otherwise the peak may be due to a later TMD. The distance to the peak was compared to the distance between the first TMD and the second (as determined by TMHMM) or the distance between the SS and the first TMD (as determined by Phobius). Peaks were considered unambiguous only if no more than 8 residues of the next TMD were translated. This value is the length of the shortest functional signal sequence in our set, controlling for any affect that an additional signal within the exit tunnel may have on SRP.

An alternative approach was used to determine codons with significant recruitment of SRP in mitochondrial transcripts and transcripts lacking an ER-targeting signal. For every ORF, a count matrix was built with codon position as rows, and replicates of SRP-bound polysome ribosome-protected reads, total soluble polysome reads and total membrane polysome reads as columns. These matrices were individually input to DESeq2 (ref. 48), and a linear model was fit using the presence or absence of SRP co-immunoprecipitation as the coefficient; in this application, codons were treated as 'genes'. Reads at each codon were used for fitting local dispersion trends. Genes with at least one codon that had at least threefold enrichment with $P < 0.001$ were selected. We note that all six SRP subunits had local enrichment towards the C terminus; since this may be cotranslational particle assembly, we omitted them from further analysis. Thirteen other genes were identified, and binding sites were assigned to the first significantly enriched codon preceding the position with maximum enrichment.

Quantification of early SRP enrichment. Secretory protein ORFs which were used in the analysis of SRP recruitment to the nascent chain (see previous section) were also tested for SRP pre-enrichment by comparing ribosome-protected reads from SRP-bound polysomes and monosomes from a CHX-treated culture. The RPM values from codon 10 to the position of the first SS or TMD, plus 40, were added, and the monosome sum was divided by the polysome sum. The first 9 sense codons were omitted to avoid artefacts near the start of transcripts. Ratios of greater than 1 were designated pre-recruited.

The enrichment in Fig. 3d was determined by first smoothing SRP-bound monosome and polysome reads (in RPM) using an 11 codon window, and then dividing the monosome values by the polysome values.

GFP reporter constructs. Sequences of the yeast *TUB2* 5′ UTR (300 nucleotides preceding the start codon), the *TUB2* 3′ UTR (300 nucleotides following the stop codon), the *PMP1* 3′ UTR (600 nucleotides), and the *PMP2* 3′ UTR (500 nucleotides) were PCR amplified from BY4741 genomic DNA with flanking overlaps to the M13 (−20) forward or M13 reverse sequences, and to the beginning or end of the sfGFP ORF sequence[52]. The sequence of sfGFP was amplified from a pET33b-derived expression vector provided by W. Clemons. Plasmids were assembled in a single reaction using Gibson assembly[53] using the M13 (−20) forward and M13 reverse sequences to amplify pRS315. Plasmids were transformed into BY4741 Srp72-TAP::HIS3MX.

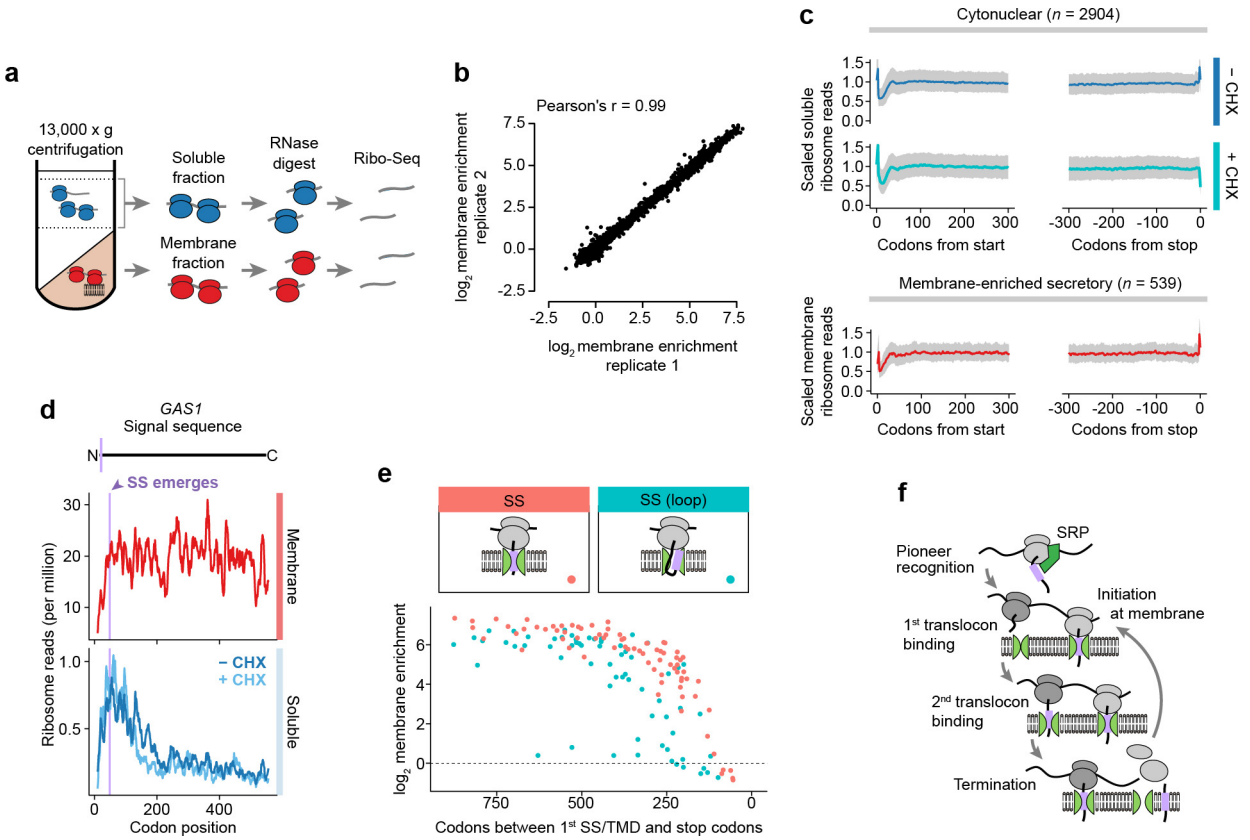
qPCR. For each biological replicate, 500 ml of synthetic complete media lacking leucine were inoculated with an overnight culture to $OD_{600\,nm}$ of 0.05. Cultures were grown at 30 °C to $OD_{600\,nm}$ 0.8–1.0 and then collected by fast filtration. Cells were lysed and fractionated, and Srp72p was immunoprecipitated as described for RNA-seq. Purified RNA were subjected to TURBO DNase digestion (Ambion). Concentrations were determined using $A_{260\,nm}$, and 100 ng of RNA was used to synthesize cDNA using iScript (Bio-Rad). qPCR was performed on a CFX-96 thermocycler using iTAQ Universal SYBR Green Supermix (Bio-Rad). *ACT1* was used as a reference for each fraction from the same culture, and enrichments were determined as fold difference of mRNA in the membrane or SRP-bound fractions over the soluble fraction. Enrichment scores from three technical replicates of the qPCR step were averaged, and biological replicates are shown.

Statistical hypothesis testing. All analysis was performed using the R programming language (https://www.r-project.org). Statistical significance in comparing distributions of SRP or membrane enrichment scores from ribosome profiling (Figs 1a and 2b), as well as in comparing hydrophobicity scores (Extended Data Fig. 2e), codon usage, or residue abundance (Extended Data Fig. 3d, e) was determined using two-sided Wilcoxon rank-sum tests. This test assumes independence of observations and does not require a normal distribution. Enrichment distributions are multimodal and so mean and variance estimates are not provided. Significance of SRP or membrane enrichment from qPCR (Fig. 3g, h and Extended Data Fig. 6c, h) was determined using a two-sided Welch's *t*-test on log-transformed enrichment values. This test assumes normal distributions but allows unequal variance. In all tests, the null hypothesis is the distributions of tested populations are equal.

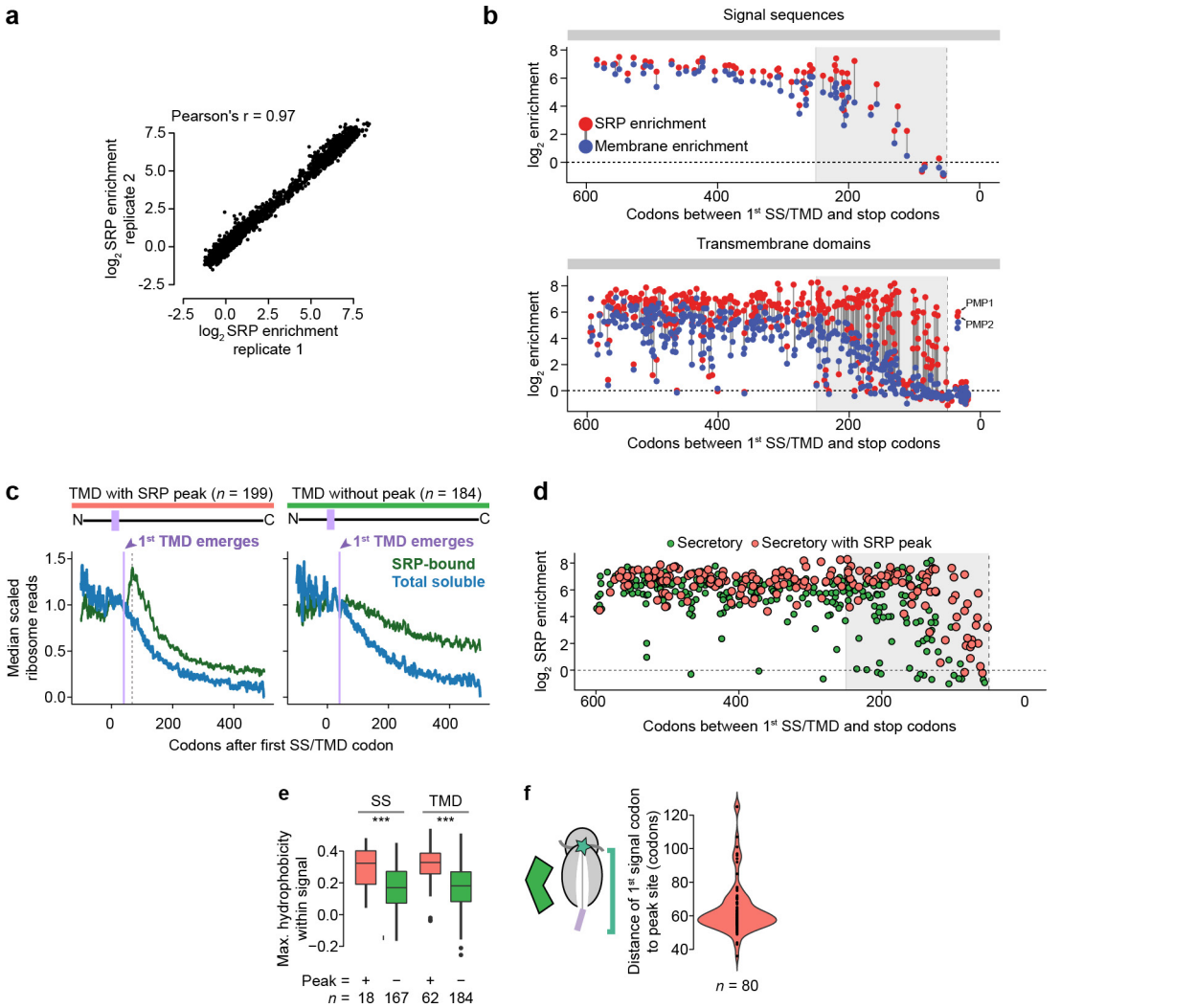
Code availability. Scripts for data processing and analysis in R are available upon request.

31. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425,** 737–741 (2003).
32. Huh, W. K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425,** 686–691 (2003).
33. Zhong, T. & Arndt, K. T. The yeast SIS1 protein, a DnaJ homolog, is required for the initiation of translation. *Cell* **73,** 1175–1186 (1993).
34. Winzeler, E. A. *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285,** 901–906 (1999).

35. Freidlin, P. J. & Patterson, R. J. Heparin releases monosomes and polysomes from rough endoplasmic reticulum. *Biochem. Biophys. Res. Commun.* **93,** 521–527 (1980).
36. Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M. & Weissman, J. S. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protocols* **7,** 1534–1550 (2012).
37. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10 (2011).
38. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
39. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).
40. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLOS Comput. Biol.* **9,** e1003118 (2013).
41. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12,** 115–121 (2015).
42. Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340,** 783–795 (2004).
43. Käll, L., Krogh, A. & Sonnhammer, E. L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **338,** 1027–1036 (2004).
44. Reynolds, S. M., Käll, L., Riffle, M. E., Bilmes, J. A. & Noble, W. S. Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLOS Comput. Biol.* **4,** e1000213 (2008).
45. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305,** 567–580 (2001).
46. Breker, M., Gymrek, M., Moldavski, O. & Schuldiner, M. LoQAtE–Localization and Quantitation ATlas of the yeast proteomE. A new tool for multiparametric dissection of single-protein behavior in response to biological perturbations in yeast. *Nucleic Acids Res.* **42,** D726–D730 (2014).
47. Becker, A. H., Oh, E., Weissman, J. S., Kramer, G. & Bukau, B. Selective ribosome profiling as a tool for studying the interaction of chaperones and targeting factors with nascent polypeptide chains and ribosomes. *Nat. Protocols* **8,** 2212–2239 (2013).
48. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).
49. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
50. Beckert, B. *et al.* Translational arrest by a prokaryotic signal recognition particle is mediated by RNA interactions. *Nat. Struct. Mol. Biol.* **22,** 767–773 (2015).
51. Murtagh, F. & Legendre, P. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *J. Classif.* **31,** 274–295 (2014).
52. Pédelacq, J. D., Cabantous, S., Tran, T., Terwilliger, T. C. & Waldo, G. S. Engineering and characterization of a superfolder green fluorescent protein. *Nat. Biotechnol.* **24,** 79–88 (2006).
53. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6,** 343–345 (2009).
54. Hessa, T. *et al.* Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* **433,** 377–381 (2005).
55. Pechmann, S. & Frydman, J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. Mol. Biol.* **20,** 237–243 (2013).
56. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147,** 789–802 (2011).
57. Doerfel, L. K. & Rodnina, M. V. Elongation factor P: Function and effects on bacterial fitness. *Biopolymers* **99,** 837–845 (2013).

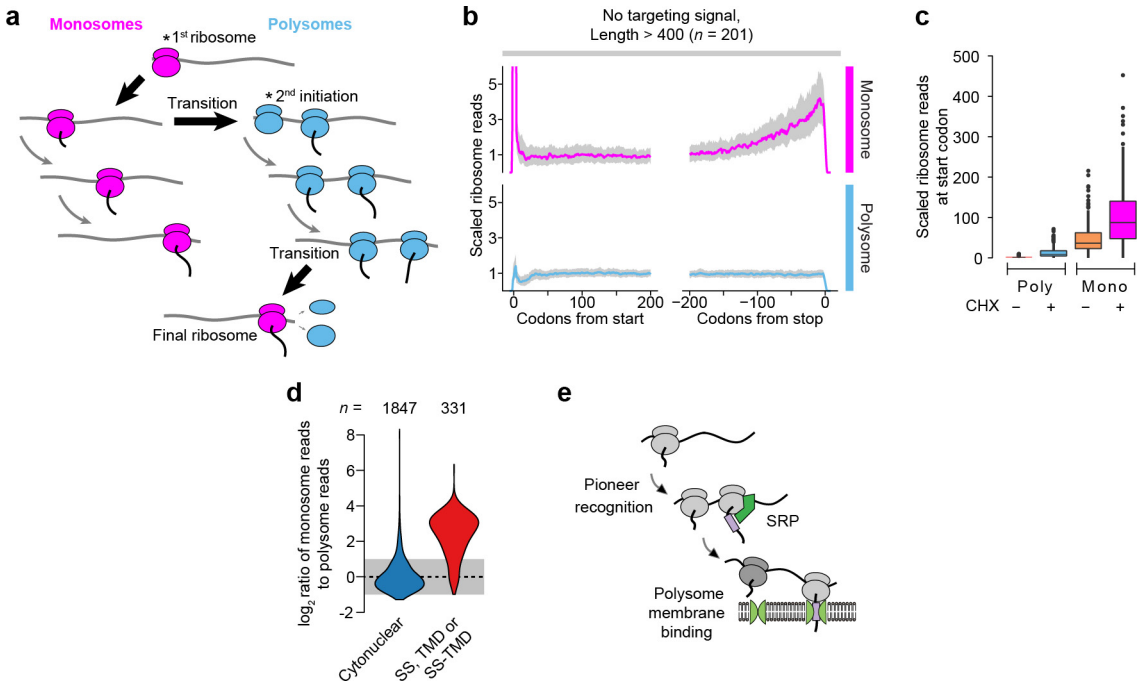


Extended Data Figure 1 | Cotranslational membrane enrichment. **a**, Crude lysates were fractionated, and then polysomes were recovered by sucrose gradient ultracentrifugation and used for ribosome profiling. **b**, Enrichment of ribosome-protected mRNA reads in the membrane polysome fractions over the soluble polysome fractions from two biological replicates. Every dot represents one ORF. **c**, Metagene plots of soluble polysome ribosome-protected reads of transcripts encoding proteins lacking ER-targeting signals (top), or of membrane-bound polysome-protected reads of transcripts encoding secretory proteins that were at least twofold membrane-enriched (bottom). For each ORF, ribosome-protected reads at each position were scaled by dividing by the mean reads per codon of the ORF, excluding the first two and last two sense codons. The median scaled reads at each position are plotted as a line, and the interquartile range is shaded in grey. **d**, Ribosome-protected reads at each codon of an example secreted protein, β-1,3-glucanosyltransferase (*GAS1*), a model SRP-independent protein[12]. Topology is indicated above, with the signal sequence in lavender. The position where the signal begins to emerge from the ribosome exit tunnel is indicated. **e**, The number of codons remaining after the encoding of the first residue of an SS, and the corresponding membrane enrichment per SS-containing ORF. Signal sequences were divided between those that bind Ssh1p directly upon exposure and those that require a looped conformation (>90 codons after the first SS codon)[16]. **f**, Transcripts remain at the membrane by subsequent translocon binding, thus the small soluble fraction comprises mRNA undergoing initial targeting.

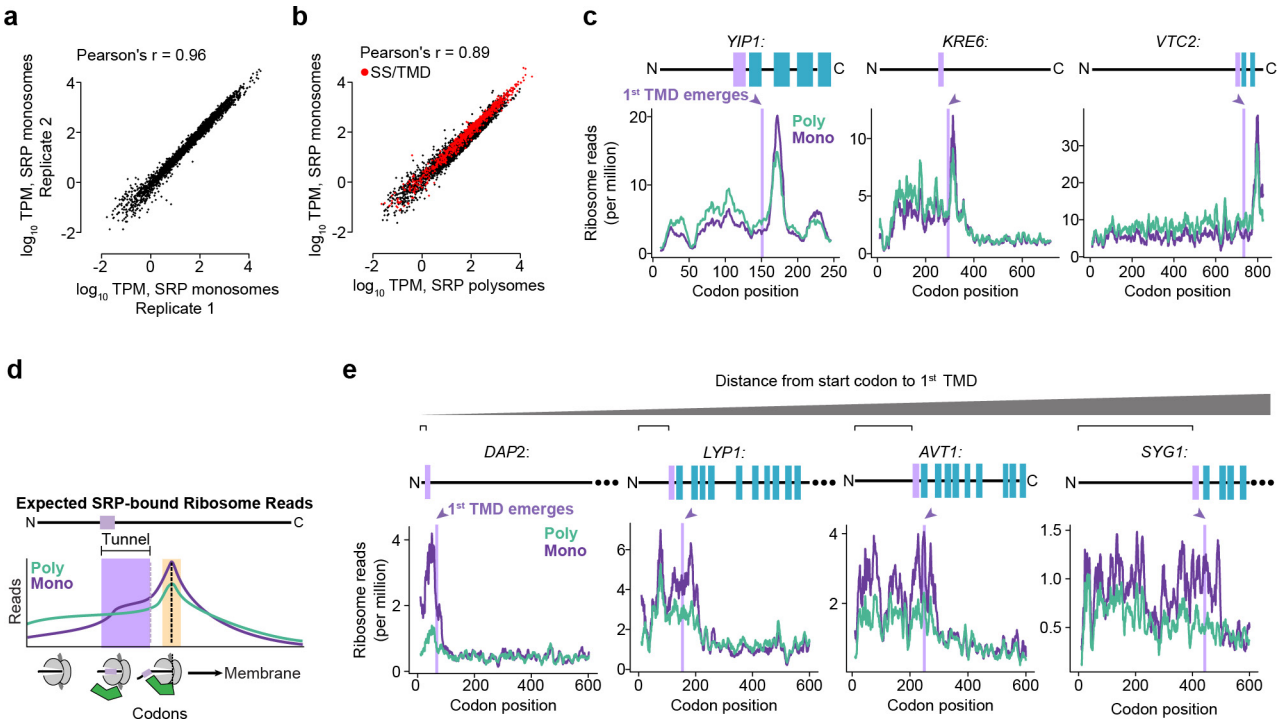


Extended Data Figure 2 | Cotranslational enrichment of SRP. **a**, Enrichment of ribosome-protected mRNA reads in the soluble SRP-bound polysome fractions over the total soluble polysome fractions from two biological replicates. **b**, The number of codons remaining after encoding of first SS or TMD residue, and the corresponding SRP and membrane enrichment scores per ORF. Scores are determined from cultures harvested without added CHX. Enrichment scores are indicated with filled dots, and the scores from the same transcript are linked with a grey line. The vertical dashed line indicates 50 codons, the boundary for tail-anchored proteins. Here, only SSs that bind Ssh1p directly after exposure from the RNC are shown. **c**, Secretory transcripts were classified into two groups based on the ribosome-protected-read distributions from SRP-bound polysomes. Some showed a pronounced increase in reads at positions coincident with the initial exposure of an SS or TMD by the ribosome, whereas others did not. Shown here are metagene analysis plots of soluble polysome-protected reads from the categorized TMD proteins. For each ORF, the reads at each codon position were divided by the mean reads per codon within the range +20 to +40 after the first signal codon. The first 30 codons of each ORF are excluded to avoid the characteristic low-density region near the start codon. The lavender line indicates when the first TMD begins to emerge from the exit tunnel, and the dashed line indicates the position of the read peak. Notably, the total soluble polysome reads depleted in a similar manner for both classes, a read increase was not observed in the total soluble reads, and reads from the SRP-bound transcripts with a peak did not deplete faster than the total soluble reads. These features are consistent with a model in which SRP is recruited at the peak site, and elongation then proceeds at the same rate. **d**, The number of codons remaining after encoding of the first SS or TMD and corresponding SRP enrichment. Transcripts are classified by the presence or absence of a read increase following signal exposure, as in **c**. Note that for SRP-enriched transcripts with signals closest to the terminus (<100 codons), evidence of direct binding between SRP and the nascent chain was always observed. SRP can therefore bind late TMDs immediately after they become exposed by the ribosome. **e**, Maximum hydrophobicity across targeting signals using an 8-residue averaging window. Only signals with peaks that could be unambiguously attributed to a targeting signal were included. Hydrophobicity was determined by attributing the biological hydrophobicity score to each encoded amino acid[54]. ***$P \leq 0.001$, Wilcoxon rank-sum test. **f**, Distribution of the distance between the first codon of a targeting signal and the position of the downstream read increase. Only transcripts wherein the increase can be unambiguously attributed to a specific targeting signal were included.

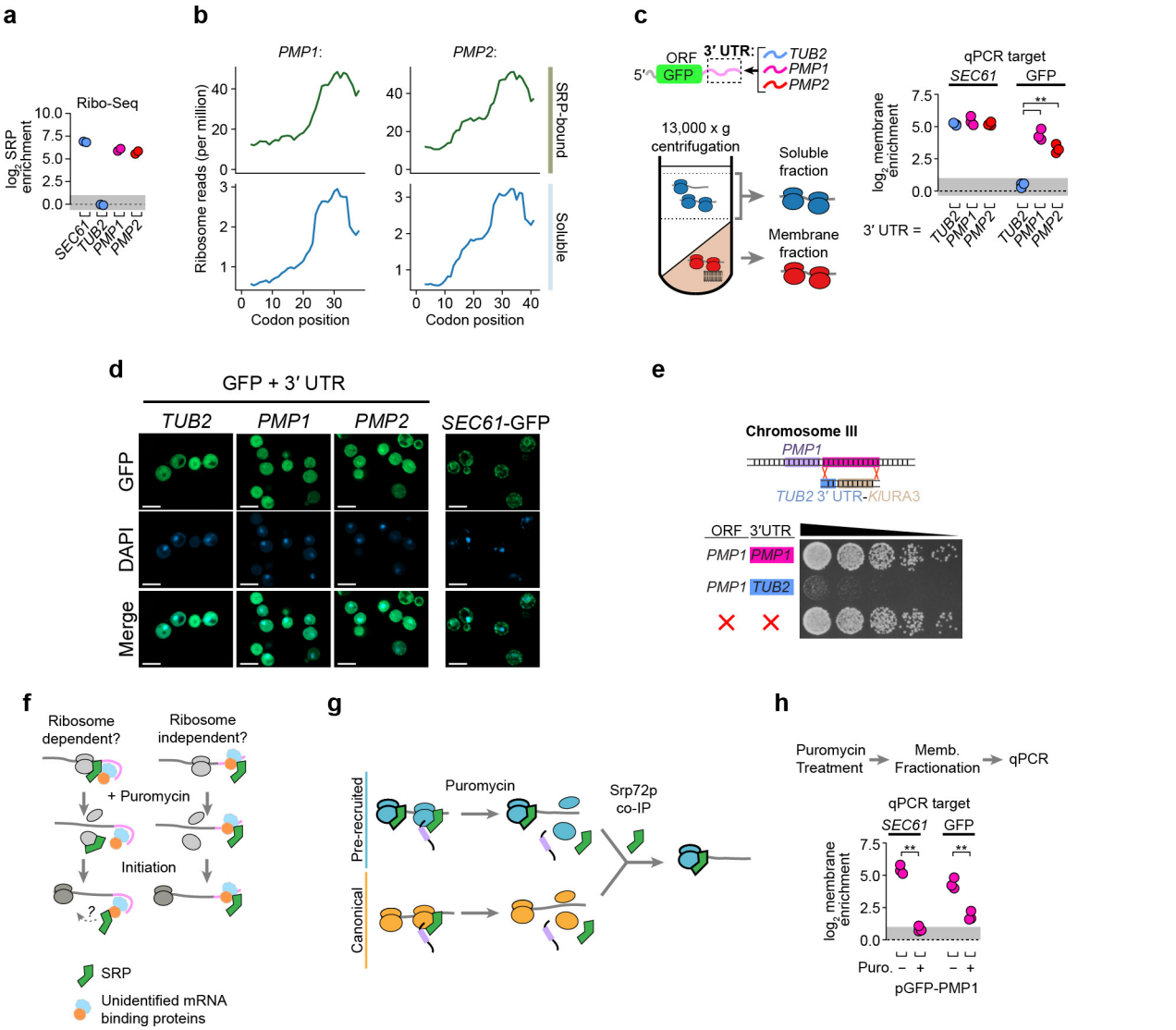
Extended Data Figure 3 | Elongation pausing and local SRP recruitment. a, **b**, Local increases in ribosome-protected reads from membrane-bound polysomes, indicated by orange lines, were coincident with rare codons, as in cell division cycle protein 1 (*CDC1*, **a**) or polybasic nascent chains, as in the plasma membrane G-protein-coupled receptor (*GPR1*, **b**). Soluble SRP-bound polysome-protected reads were further increased at the same positions. **c**, In these cases, hydrophobic sequences in the nascent chain were exposed to the cytosol at the locations of increased reads, which were coincident with elongation attenuators. **d**, Translational efficiencies for the 6 codons following, and the number of stalling residues within the 10 residues preceding, the sites of increased SRP-bound ribosome reads. Translational efficiency was determined by attributing the normalized translational efficiency (nTE) score to each codon[55]. Residues that were found to stall the ribosome, based on previous investigation[20,56,57], were lysine, arginine, glutamate, aspartate, proline and glycine. Because of variation in specific motifs, and uncertainty in whether these motifs are additive, we simply compared the total number of these residues in the indicated 10 residue spans. Sets of 10,000 random sequences, at least 10 amino acids from the stop codon, were sampled from 5,907 non-dubious ORFs, and translational efficiency and stalling residues were determined over 6 or 10 codon spans. $^{*}P \leq 0.05$, $^{**}P \leq 0.01$, Wilcoxon rank-sum tests. **e**, The targeting signals that recruited SRP directly to the nascent chain unusually far from the encoding of the signal had SRP-binding sites coincident with intrinsic elongation attenuation. Secretory protein transcripts that showed an increase in SRP-bound protected reads (see Extended Data Fig. 2c, f) were further classified by the position of the peak relative to the first signal codon. Transcripts with peaks found at least 80 codons after the signal had significantly lower translational efficiency in the 6 codons following the peak. These transcripts also had a greater, but not statistically significant, amount of stalling amino acids in the 10 residues preceding the peak. $^{*}P \leq 0.05$, Wilcoxon rank-sum tests. **f**, Similar increases in SRP-bound reads were observed for certain non-secretory proteins as exemplified by phosphoacetylglucosamine mutase (*PCM1*) and tRNA^{Ser} Um_{44} 2′-O-methyltransferase (*TRM44*). Hydrophobic sequences in non-secretory proteins, coupled with attenuation of elongation, may lead to SRP recruitment.



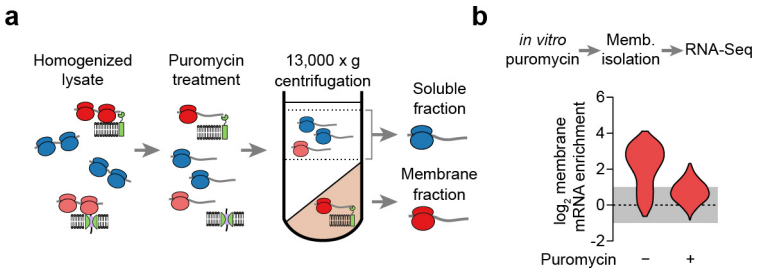
Extended Data Figure 4 | Ribosome profiling of monosomes. **a**, Ribosomes transition from monosomes to polysomes during elongation. The pioneer round of initiation will be a monosome, and during elongation there is a chance of additional initiation converting the transcript to a polysome. Similarly, a polysome will become a monosome if all ribosomes but one terminate. As mRNA is sampled closer to the stop codon, the likelihood of observing a footprint from the final ribosome will increase. **b**, Metagene analysis of soluble monosome- or polysome-protected reads from proteins lacking an ER-targeting signal. Data were obtained using CHX treatment. ORFs are at least 400 codons long and have an average of at least 0.5 reads per codon in each data set. For each ORF, ribosome reads at each position were divided by the mean reads per codon over the range +160 to +240 codons. The median normalized read value at each codon position is plotted, and the interquartile range is shaded in grey. **c**, Relative reads at the start codon from ORFs normalized in **b**. **d**, Distributions of the ratio of ribosome-protected reads found in soluble monosomes over soluble polysomes. **e**, A pioneer round of translation deposits mRNA on the membrane. Polysomes will be retained at the membrane and are therefore depleted from the soluble fraction.



Extended Data Figure 5 | Ribosome profiling of SRP-bound monosomes. a, Ribosome-protected reads, in tags per million (TPM) for each ORF, from SRP-bound monosome fractions from two biological replicates. **b**, Ribosome-protected reads from the soluble SRP-bound monosome and SRP-bound polysome fractions of the same biological replicate, with CHX treatment. **c**, Distribution of ribosome reads within example ORFs that display SRP-bound monosome and polysome profiles consistent with direct recognition of the nascent chain. **d**, If RNCs can recruit SRP while a TMD is within the exit tunnel, then there will be an increase in ribosome-protected reads from SRP-bound monosomes when the TMD begins to translate (lavender). This increase will maximize when the TMD is exposed to the cytosol (orange). **e**, Distribution of ribosome reads within example ORFs that display SRP-bound monosome profiles consistent with recruitment to transcripts before targeting signal synthesis. Examples are arranged for an increasing distance from the start codon to the first TMD. Only the first 600 codons for each ORF are shown.



Extended Data Figure 6 | The role of the UTR from *PMP1* and *PMP2*. **a**, The cotranslational SRP enrichment of the *PMP1* and *PMP2* ORFs was similar to other bona fide secretory proteins, such as *SEC61*. By contrast, cytosolic proteins such as tubulin (*TUB2*) were not enriched. The enrichment scores are determined from the SRP-bound and total soluble polysomes from two biological replicates collected without added CHX. **b**, Distribution of ribosome-protected reads from soluble polysomes within the *PMP1* and *PMP2* ORFs. **c**, Membrane enrichment, determined by qPCR, of the mRNA of GFP fused to the indicated 3′ UTRs. The coding sequence of endogenous *SEC61* transcript was also amplified as a control for a membrane-localized transcript. $^{**}P \leq 0.01$, $n = 3$ biological replicates, Welch's *t*-test. **d**, Localization of mature GFP. Scale bar, 5 μm. Yeast were grown to mid-log phase and imaged using an Axio Observer Z1 with a Plan-Apochromat 100 × /1.4 oil immersion objective (Zeiss). *Z*-stacks were deconvoluted by the iterative maximum likelihood algorithm in ZEN (Zeiss) and single planes are shown. Images were representative from a set of two replicated assays. **e**, Yeast growth after replacement of the endogenous 3′ UTR of *PMP1* with the 3′ UTR of tubulin. Also shown is a complete deletion of *PMP1* ORF[34]. Gibson assembly[53] was used to fuse the 300-nucleotide *TUB2* 3′ UTR to the *Kl*URA3 cassette into SmaI digested pUC19. The TUB2-UTR-URA3 element was PCR amplified, including 40-nucleotide overhangs matching genomic sequences, and replaced the 650 nucleotides immediately following the *PMP1* coding sequence in strain BY4741 by homologous recombination. Image is representative from a set of 3 replicated assays. **f**, Nascent-chain-independent SRP recognition may require ribosomes. Puromycin treatment of lysates disrupts elongating, but not initiating, ribosomes. **g**, Transcripts showing only canonical recognition are more sensitive to puromycin. This is consistent with puromycin resistance of SRP that has pre-recruited to initiating ribosomes. **h**, Membrane enrichment of the GFP-PMP1 construct or *SEC61* mRNA after lysates were incubated with puromycin. $^{**}P \leq 0.01$, $n = 3$ biological replicates, Welch's *t*-test.



Extended Data Figure 7 | The role translation in membrane enrichment. a, Lysates were treated with puromycin before membrane fractionation. mRNA recovered from the soluble and membrane fractions were used for RNA-seq **b**, Membrane enrichment of secretory protein transcripts (SS, TMD, SS-TMD, or TA, $n = 729$) following puromycin treatment of lysates.

LETTER

doi:10.1038/nature19079

Mechanism of arginine sensing by CASTOR1 upstream of mTORC1

Robert A. Saxton[1,2,3,4,5], Lynne Chantranupong[1,2,3,4,5], Kevin E. Knockenhauer[1], Thomas U. Schwartz[1] & David M. Sabatini[1,2,3,4,5]

The mechanistic Target of Rapamycin Complex 1 (mTORC1) is a major regulator of eukaryotic growth that coordinates anabolic and catabolic cellular processes with inputs such as growth factors and nutrients, including amino acids[1–3]. In mammals arginine is particularly important, promoting diverse physiological effects such as immune cell activation, insulin secretion, and muscle growth, largely mediated through activation of mTORC1 (refs 4–7). Arginine activates mTORC1 upstream of the Rag family of GTPases[8], through either the lysosomal amino acid transporter SLC38A9 or the GATOR2-interacting Cellular Arginine Sensor for mTORC1 (CASTOR1)[9–12]. However, the mechanism by which the mTORC1 pathway detects and transmits this arginine signal has been elusive. Here, we present the 1.8 Å crystal structure of arginine-bound CASTOR1. Homodimeric CASTOR1 binds arginine at the interface of two Aspartate kinase, Chorismate mutase, TyrA (ACT) domains, enabling allosteric control of the adjacent GATOR2-binding site to trigger dissociation from GATOR2 and downstream activation of mTORC1. Our data reveal that CASTOR1 shares substantial structural homology with the lysine-binding regulatory domain of prokaryotic aspartate kinases, suggesting that the mTORC1 pathway exploited an ancient, amino-acid-dependent allosteric mechanism to acquire arginine sensitivity. Together, these results establish a structural basis for arginine sensing by the mTORC1 pathway and provide insights into the evolution of a mammalian nutrient sensor.

To understand the molecular mechanisms through which CASTOR1 detects the presence of arginine and signals it to mTORC1, we determined the crystal structure of arginine-bound CASTOR1 to 1.8 Å resolution (Extended Data Table 1). Our findings show that CASTOR1 forms a rod-shaped homodimer, with the monomers associated in a side-by-side manner and rotated 180° with respect to each other (Fig. 1a). Although sequence analysis of CASTOR1 predicted the presence of two ACT domains[12,13], the structure reveals that each monomer actually contains four tandem ACT domains. ACT1 displays the canonical βαββαβ ACT domain topology[14,15], whereas ACT2 contains two additional β-strands and ACT3 and ACT4 each lack the final β-strand. (Fig. 1a and Extended Data Fig. 1a).

The dimerization interface buries around 950 Å^2 of surface area at the intersection between the α1 helix of ACT1 and the α5 helix of ACT3 (Fig. 1b). Two inward-facing isoleucine residues of each monomer (Ile28 and Ile202) form the hydrophobic core of the symmetrical interface, flanked on each side by tyrosine–histidine pairs (His25 and Tyr207) that form both π-stacking and hydrogen-bond contacts with the opposing monomer (Fig. 1b). To understand the importance of dimerization in CASTOR1 function, we generated constitutively monomeric mutants of CASTOR1 (Y207S and I202E; Fig. 1c). Notably, although dimerization is dispensable for arginine binding (Extended Data Fig. 2a), these mutants interacted weakly with GATOR2 and failed to inhibit mTORC1 signalling in cells (Fig. 1c and Extended Data Fig. 2b). This finding indicates that CASTOR1 must be dimeric to robustly inhibit GATOR2 upon arginine starvation.

CASTOR1 binds arginine through a narrow pocket at the interface of ACT2 and ACT4, distal to the dimerization interface (Fig. 1a, 2a, b). The side chain of arginine projects towards the β15 loop, a loop connecting β15 and β16, where the backbone carbonyls of Thr300, Phe301, and Phe303 coordinate the guanidinium group of arginine (Fig. 2a). Immediately adjacent to the β15 loop, the anionic side chain of Asp304 forms an additional stabilizing salt bridge with the cationic arginine side chain (Fig. 2a). On the opposite side of the pocket, the hydroxyl side chain of Ser111 and the backbone carbonyl of Val112 in the α3 loop anchor the free amino group of arginine in place, while the free carboxyl group points towards a water-filled cavity that separates it from ACT2 (Fig. 2a, b). Mutation of either Ser111 or Asp304 (S111A, D304A) abolished the arginine-binding ability of CASTOR1 *in vitro*, highlighting the critical role of these contacts in arginine sensing by CASTOR1 (Fig. 2c). Furthermore, when expressed in HEK-293T cells, these mutants bound constitutively to GATOR2 and strongly inhibited mTORC1 signalling even in the presence of arginine (Fig. 2d).

Together, these data explain the molecular determinants of specificity in the CASTOR1–arginine interaction. While Ser111 fixes the position of the free amine, the location of the β15 loop and Asp304 sets a strict length requirement for the bound ligand (Extended Data Fig. 3a). In addition, the positions of the three hydrogen-bond-donating nitrogen atoms in the guanidinium group facilitate contacts with both the carbonyl oxygen atoms in the β15 loop and the side chain of Asp304 (Fig. 2a). Finally, the gap behind the free carboxyl group of arginine suggests that CASTOR1 can tolerate ligands with modifications to that functional group (Fig. 2b). We tested these predictions by investigating the ability of various arginine analogues to disrupt the CASTOR1–GATOR2 interaction *in vitro* (Fig. 2e and Extended Data Fig. 3b). Consistent with our structural analysis, while the carboxy-modified arginine-methyl ester triggered full dissociation of CASTOR1 from GATOR2, compounds with alterations to the guanidinium group, α-amine, or the length of the side chain had no effect.

In addition to the main pocket contacts described above, a highly conserved, glycine-rich loop connecting β14 and α7 in ACT4 (β14 loop, residues 269–280) wraps over the arginine pocket, fully burying the bound ligand (Figs 2a, 3a and Extended Data Fig. 1a). The β14 loop forms several hydrogen bonds with arginine through the backbone amides of Gly279 and Ile280, as well as the backbone oxygen atoms of Gly274 and Glu277 (Figs 2a, 3a). The ordered conformation of the β14 loop also places it just along the ACT2–ACT4 interface, enabling it to form several intramolecular contacts with residues in ACT2 (Fig. 3a). Cys278 forms hydrogen bonds with the backbones of Val110 and S111 in the α3 loop, while Asp276 forms a salt bridge with Arg126. In addition, Glu277 extends in the opposite direction to form another salt bridge with His175 (Fig. 3a). Thus, the β14 loop facilitates

[1]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [2]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA. [3]Howard Hughes Medical Institute, Cambridge, Massachusetts 02139, USA. [4]Koch Institute for Integrative Cancer Research, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA. [5]Broad Institute of Harvard and Massachusetts Institute of Technology, 415 Main Street, Cambridge, Massachusetts 02142, USA.

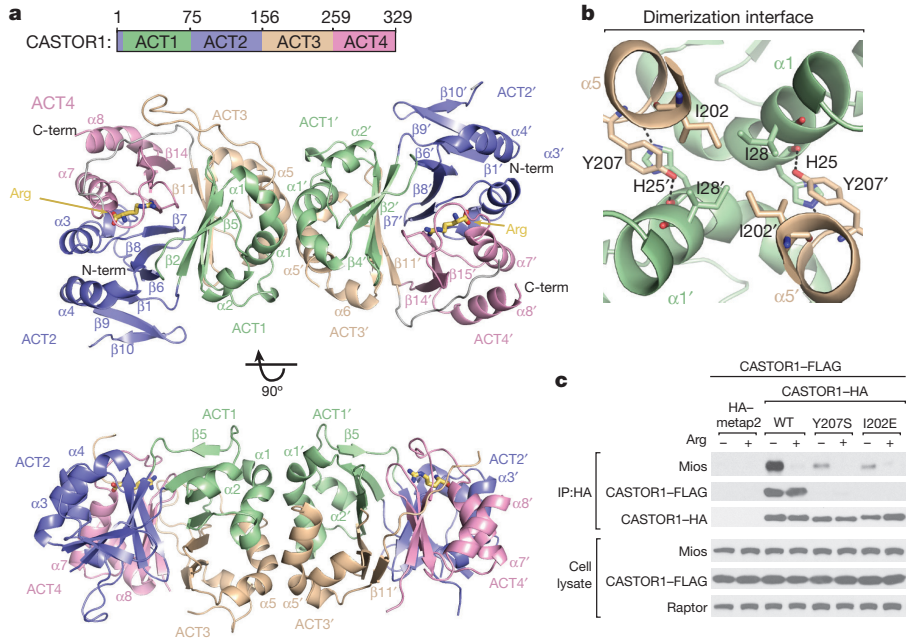


Figure 1 | Architecture of human CASTOR1. **a**, Two orthogonal views of the CASTOR1 homodimer (ribbon diagram), with ACT-domains 1–4 coloured in green, purple, wheat, and pink, respectively. The bound arginine is shown in yellow. Disordered regions not observed in the crystal structure are omitted. **b**, View of the CASTOR1 dimerization interface, with side chains of key residues represented in stick form. **c**, Dimerization-deficient CASTOR1 Y207S and I202E mutants display weaker interactions with endogenous GATOR2. HEK-293T cells transiently expressing FLAG-tagged CASTOR1 wild type (WT) and the indicated haemagglutinin (HA)-tagged constructs were starved of arginine for 50 min and, where indicated, re-stimulated for 10 min. HA-immunoprecipitates were generated from cell lysates and analysed by immunoblotting for the indicated proteins. Mios was used as a representative GATOR2 component.

the formation of numerous inter-ACT-domain contacts in the presence of arginine. Indeed, the arginine and β14 loop contribute about 40% of the total buried surface area in the ACT2–ACT4 interface of the arginine-bound structure (390 Å^2 out of 980 Å^2).

The glycine-rich β14 loop is predicted to have a high propensity for disorder. Our structure suggests that these inter-ACT-domain contacts could stabilize it in an ordered conformation over the bound arginine. Indeed, mutation of key residues in both the β14 loop (D276A, E277A, C278A) and the adjacent ACT domains (R126A, H175A) significantly reduced the arginine-binding capacity of CASTOR1 (Fig. 3b, c), indicating that the inter-ACT-domain contacts formed by the β14 loop are required for arginine sensing by CASTOR1. In addition, we found that the N-terminal (ACT1 and ACT2) and C-terminal (ACT3 and ACT4) halves of CASTOR1 associated in both an arginine- and

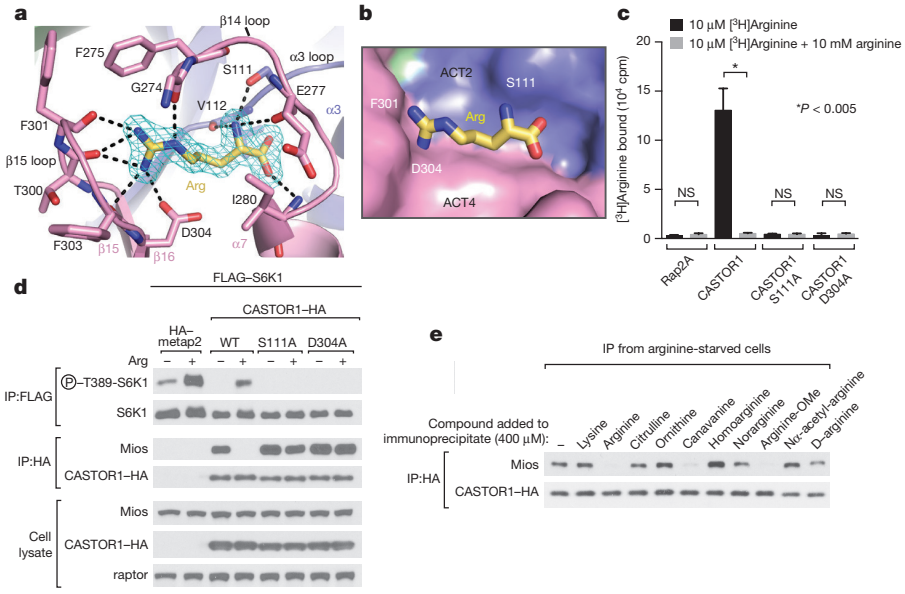


Figure 2 | The arginine-binding pocket of CASTOR1. **a**, View of the arginine-binding pocket in CASTOR1, together with its F_o-F_c electron density map calculated and contoured at 4σ from an omit map lacking arginine. The bound arginine is shown in yellow. Hydrogen bonds or salt bridges are shown as black dashed lines. Residues 269–273 are omitted for clarity. **b**, Steric view of the arginine-binding pocket, depicting the surface representation of CASTOR1 and stick model of arginine (yellow). The β14 loop (residues 269–280) is omitted for clarity. **c**, CASTOR1 S111A and D304A mutants do not bind arginine *in vitro*. FLAG-immunoprecipitates prepared from HEK-293T cells transiently expressing the indicated FLAG-tagged proteins were used in binding assays with [^{3}H]arginine as described in the Methods. Values are mean ± s.d. for three technical replicates from one representative experiment. **d**, The CASTOR1 S111A and D304A mutants constitutively bind GATOR2 and inhibit mTORC1 signalling in cells. HEK-293T cells transiently expressing FLAG–S6K1 and the indicated HA-tagged constructs were starved of arginine for 50 min and, where indicated, re-stimulated for 10 min. Both FLAG- and HA-immunoprecipitates were prepared from lysates and analysed as in Fig. 1c. **e**, Effects of various arginine analogues on the CASTOR1–GATOR2 interaction *in vitro*. HEK-293T cells transiently expressing wild-type HA–CASTOR1 were starved of arginine for 50 min. HA-immunoprecipitates were prepared from cell lysates then incubated with 400 μM of the indicated compounds for 20 min and analysed as in Fig. 1c.

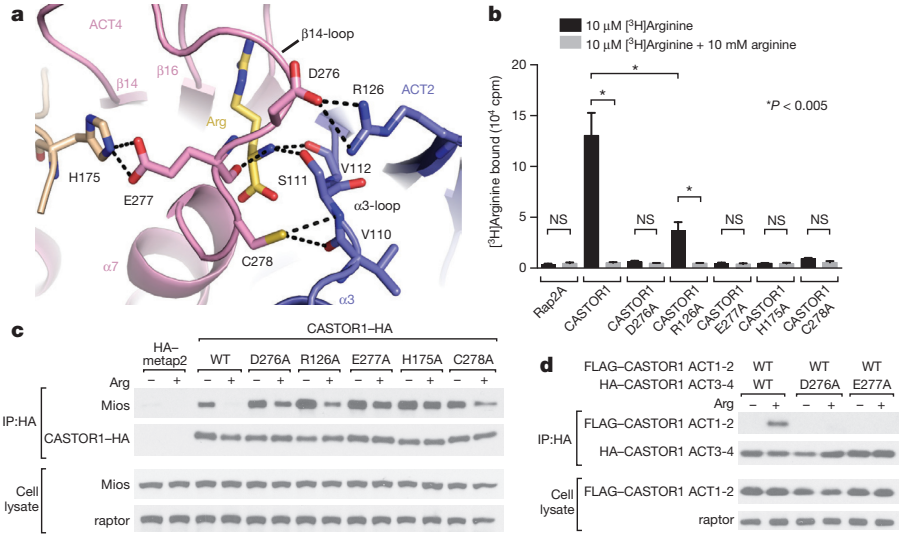


Figure 3 | Arginine facilitates the intramolecular association of the ACT2 and ACT4 domains of CASTOR1. a, Top-down view of the arginine- and β14-loop-mediated contacts between ACT2 and ACT4. Hydrogen bonds and salt bridges are shown as black dashed lines. **b**, The CASTOR1 D276A, R126A, E277A, H175A, and C278A mutants display reduced arginine-binding capacity *in vitro*. Binding assays were performed and immunoprecipitates analysed as in Fig. 2c. Values are mean ± s.d. for three technical replicates from one representative experiment. **c**, The CASTOR1 D276A, R126A, E277A, H175A, and C278A mutants constitutively bind GATOR2 in cells. HEK-293T cells transiently expressing the indicated HA-tagged constructs were starved of arginine for 50 min and, where indicated, re-stimulated for 10 min. HA-immunoprecipitates were prepared and analysed as in Fig. 1c. **d**, CASTOR1 ACT1-2 (residues 1–169) and CASTOR1 ACT3–4 (169–329) associate in an arginine- and β14-loop-dependent manner. HEK-293T cells transiently expressing the indicated HA-tagged constructs were starved of arginine for 60 min and, where indicated, re-stimulated for 60 min. HA-immunoprecipitates were prepared and analysed as in Fig. 1c.

β14-loop-dependent manner when expressed as separate polypeptides in HEK-293T cells[12] (Fig. 3d), indicating that arginine probably induces a conformational change in CASTOR1 by stabilizing the ACT2–ACT4 interaction.

In addition to CASTOR1, human cells express a related protein, CASTOR2, which shares 63% sequence identity with CASTOR1 but does not bind arginine[12]. Although the regions of CASTOR1 that are directly involved in arginine binding are well conserved (Extended Data Fig. 1a), we identified residues along the ACT2–ACT4 interface (His108 to Val110) that differ between CASTOR1 and CASTOR2 (Extended Data Fig. 4a). Replacing these residues in CASTOR1 with those from CASTOR2 abrogated arginine binding *in vitro* and converted CASTOR1 to a nearly-constitutive GATOR2-interactor in cells, resembling CASTOR2 (Extended Data Fig. 4b–d). Notably, these residues immediately precede Ser111 and form a hydrogen bond with Cys278 in the β14 loop (Fig. 3a and Extended Data Fig. 4a), suggesting that their identity may be critical for the proper positioning of the α3 loop to enable arginine binding and/or the association of ACT2 and ACT4. The corresponding mutation in CASTOR2 (QNI108–110HHV), however, was not sufficient to confer arginine-binding ability, suggesting that additional amino acid differences also contribute to this functional difference (Extended Data Fig. 4d).

To understand how arginine induces dissociation of CASTOR1 from GATOR2, we identified five highly conserved sites in CASTOR1 that are required for its interaction with GATOR2 (Y118, Q119, D121, E261 and D292; Fig. 4a and Extended Data Fig. 1a). Importantly, these mutants still bind arginine *in vitro* and homodimerize when expressed in cells (Extended Data Fig. 5a, b). Notably, these residues cluster along the surface of the ACT2–ACT4 interface, adjacent to the arginine-binding pocket but on the opposite face of the protein (Fig. 4b, c). Glu261 and Asp292 are closely linked to the β14 loop, separated only by β14 and α7, respectively (Fig. 4c). Furthermore, the critically important residue Asp121 is buried in the ACT2–ACT4 interface, potentially explaining why the arginine-bound conformation of CASTOR1 does not interact with GATOR2 (Fig. 4c).

Together, these results suggest a model in which arginine binding arranges the glycine-rich β14 loop in a conformation that enables the intramolecular association of ACT2 and ACT4 (Fig. 3a–d). The association of these domains would alter the position and exposure of the residues required for GATOR2 binding, which also lie along the ACT2–ACT4 interface (Fig. 4a–c), thereby triggering the dissociation of CASTOR1 from GATOR2 and the subsequent activation of mTORC1 (Fig. 4e).

The observation that CASTOR1 inhibits mTORC1 signalling and interacts with GATOR2 in an arginine-sensitive manner suggests that CASTOR1 may regulate mTORC1 by inhibiting GATOR2, a mechanism analogous to that of the recently identified leucine sensor Sestrin2 (refs 16–19). Using our GATOR2-binding-deficient mutants, we were able to test this hypothesis directly. In contrast to wild-type CASTOR1, the GATOR2-binding-deficient YQ118–119AA and D121A mutants both failed to inhibit mTORC1 signalling in cells (Fig. 4d). Moreover, owing to their ability to dimerize with endogenous CASTOR1, these mutants also functioned as dominant negatives, rendering mTORC1 fully resistant to arginine starvation (Fig. 4d). Thus, the CASTOR1–GATOR2 interaction is required to signal arginine deprivation to mTORC1.

Although defined by their common topology, ACT domains are highly diverse in sequence and form a wide range of structural assemblies[14,15]. Comparison of our structure with other ACT-domain-containing proteins in the Protein Data Bank (PDB) revealed that CASTOR1 shares substantial structural homology with the allosteric regulatory domains of bacterial aspartate kinases, including those found in *Escherichia coli* (AKeco) and cyanobacteria (AKsyn)[20,21] (Fig. 5a and Extended Data Fig. 6a). Aspartate kinases catalyse the first step of a metabolic pathway that synthesizes several amino acids, including lysine, and display allosteric feedback inhibition when downstream products bind to their regulatory domains[22]. Notably, AKeco binds lysine through pockets that bear a striking resemblance to the arginine-binding pocket of CASTOR1[20] (Fig. 5b). Furthermore, AKeco residues Arg305, Glu346, and Val347, which correspond to the positions of the critical GATOR2-binding residues Glu261, Tyr118, and Gln119, respectively, participate directly in the lysine-dependent inhibition of the kinase domain in AKeco[20] (Extended Data Fig. 6b). Thus, the overall structure, mode of amino-acid binding and likely allosteric

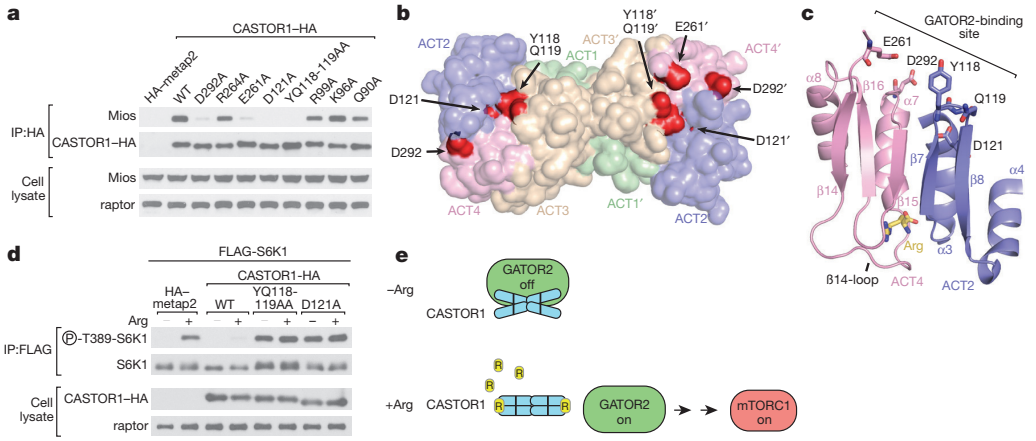


Figure 4 | The GATOR2 binding site of CASTOR1 is at the ACT2–ACT4 interface and is required for signalling arginine deprivation to mTORC1. a, The CASTOR1 D292A, E261A, D121A, and YQ118-119AA mutants are deficient in GATOR2 binding. HA-immunoprecipitates prepared from arginine-starved HEK293T-cells transiently expressing the indicated HA-tagged constructs were analysed as in Fig. 1c. **b**, Solvent-exposed surface view of the CASTOR1 homodimer highlighting the GATOR2-binding sites (red). Residue E261 is in a partially disordered loop and not visible in one monomer (left). **c**, Cross-sectional view of the ACT2–ACT4 interface showing the positions of the critical GATOR2-binding residues relative to the bound arginine (yellow) and the β14 loop. **d**, The GATOR2-binding-deficient YQ118–119AA and D121A mutants of CASTOR1 fail to inhibit the mTORC1 pathway and render cells insensitive to arginine starvation. HEK-293T cells were transiently transfected with FLAG–S6K1 and the indicated HA-tagged constructs. FLAG-immunoprecipitates were prepared and analysed as in Fig. 1d. **e**, A model of how arginine releases CASTOR1 from GATOR2 to activate mTORC1.

mechanism of CASTOR1 all resemble those found in the regulatory domains of prokaryotic aspartate kinases.

These similarities suggest that CASTOR1 shares an evolutionary origin with prokaryotic aspartate kinases. Aspartate kinases are found throughout the bacteria, archaea, and many eukaryotic lineages, but were lost before the emergence of metazoa, whereas CASTOR1 homologues are present only in metazoa (Fig. 5c). Thus, in order to acquire arginine sensitivity in early multicellular animals, the mTORC1 pathway may have taken advantage of this more ancient, lysine-sensitive regulatory mechanism (Fig. 5d). This exploitation of a pre-existing allosteric module is analogous to the models proposed for the evolution of hormone–receptor signalling[23] and yeast MAP kinases[24], and may enable the more rapid incorporation of novel signalling responses into existing pathways[25].

Together, our results provide a structural basis for arginine sensing by the mTORC1 pathway. Furthermore, our data obtained using arginine

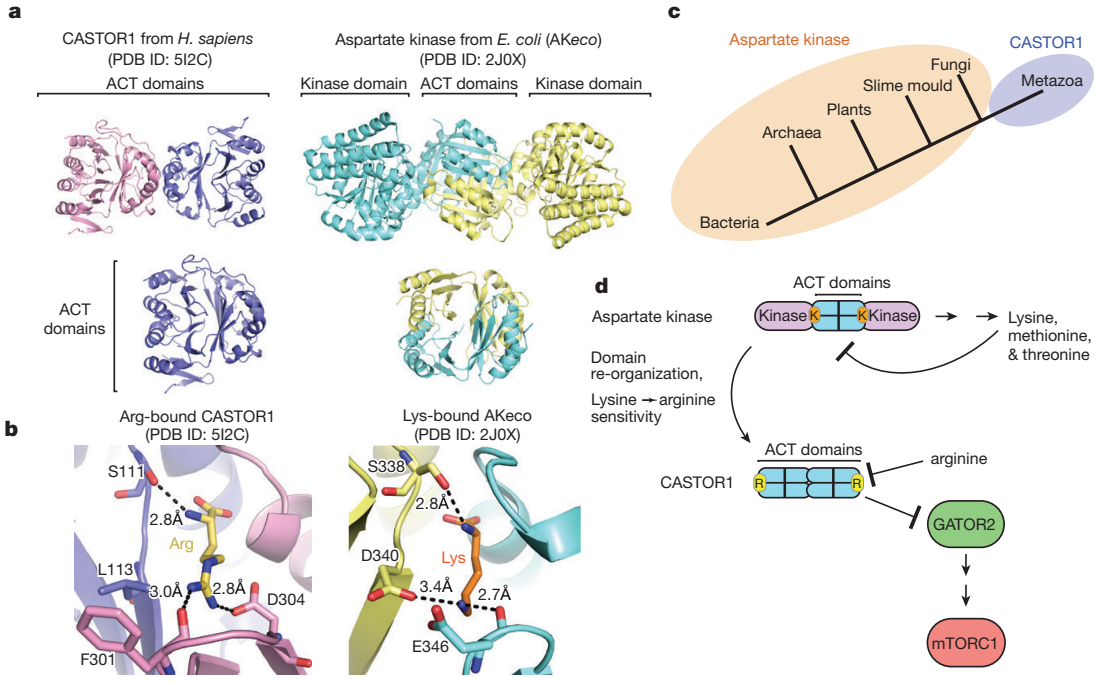


Figure 5 | Insights into the evolution of arginine sensing by CASTOR1. a, Top, a ribbon view of human CASTOR1 dimer (pink and purple) and AKeco dimer (blue and yellow; PDB ID 2J0X). Bottom, a ribbon view of the human CASTOR1 monomer (left) and the regulatory domain from AKeco (right). **b**, Comparison of the arginine-binding pocket in human CASTOR1 with the lysine-binding pocket in AKeco. Arginine and lysine are shown in yellow and orange, respectively. Hydrogen bonds and salt bridges are shown as black dashed lines. **c**, Phylogenetic distribution of aspartate kinase (orange) and CASTOR1 homologues (purple). **d**, Model of the evolution of CASTOR1 from the regulatory domain of an ancestral aspartate kinase.

analogues suggest that our structure may be useful for predicting compounds that can modulate arginine sensing by CASTOR1 *in vivo*. As the deregulation of mTORC1 is common in a number of human diseases, including cancer[26,27], the identification of novel pharmacological regulators of mTORC1 activity is of particular interest.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 10 February; accepted 5 July 2016.
Published online 3 August 2016.

1. Laplante, M. & Sabatini, D. M. mTOR signaling in growth control and disease. *Cell* **149,** 274–293 (2012).
2. Dibble, C. C. & Manning, B. D. Signal integration by mTORC1 coordinates nutrient input with biosynthetic output. *Nat. Cell Biol.* **15,** 555–564 (2013).
3. Jewell, J. L., Russell, R. C. & Guan, K. L. Amino acid signalling upstream of mTOR. *Nat. Rev. Mol. Cell Biol.* **14,** 133–139 (2013).
4. Ban, H. *et al.* Arginine and Leucine regulate p70 S6 kinase and 4E-BP1 in intestinal epithelial cells. *Int. J. Mol. Med.* **13,** 537–543 (2004).
5. Bronte, V. & Zanovello, P. Regulation of immune responses by L-arginine metabolism. *Nat. Rev. Immunol.* **5,** 641–654 (2005).
6. Floyd, J. C., Jr, Fajans, S. S., Conn, J. W., Knopf, R. F. & Rull, J. Stimulation of insulin secretion by amino acids. *J. Clin. Invest.* **45,** 1487–1502 (1966).
7. Yao, K. *et al.* Dietary arginine supplementation increases mTOR signaling activity in skeletal muscle of neonatal pigs. *J. Nutr.* **138,** 867–872 (2008).
8. Sancak, Y. *et al.* The Rag GTPases bind raptor and mediate amino acid signaling to mTORC1. *Science* **320,** 1496–1501 (2008).
9. Bar-Peled, L. *et al.* A Tumor suppressor complex with GAP activity for the Rag GTPases that signal amino acid sufficiency to mTORC1. *Science* **340,** 1100–1106 (2013).
10. Wang, S. *et al.* Metabolism. Lysosomal amino acid transporter SLC38A9 signals arginine sufficiency to mTORC1. *Science* **347,** 188–194 (2015).
11. Rebsamen, M. *et al.* SLC38A9 is a component of the lysosomal amino acid sensing machinery that controls mTORC1. *Nature* **519,** 477–481 (2015).
12. Chantranupong, L. *et al.* The CASTOR proteins are arginine sensors for the mTORC1 pathway. *Cell* **165,** 153–164 (2016).
13. Aravind, L. & Koonin, E. V. Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *J. Mol. Biol.* **287,** 1023–1040 (1999).
14. Grant, G. A. The ACT domain: a small molecule binding domain and its role as a common regulatory element. *J. Biol. Chem.* **281,** 33825–33829 (2006).
15. Chipman, D. M. & Shaanan, B. The ACT domain family. *Curr. Opin. Struct. Biol.* **11,** 694–700 (2001).
16. Chantranupong, L. *et al.* The Sestrins interact with GATOR2 to negatively regulate the amino-acid-sensing pathway upstream of mTORC1. *Cell Reports* **9,** 1–8 (2014).
17. Parmigiani, A. *et al.* Sestrins inhibit mTORC1 kinase activation through the GATOR complex. *Cell Reports* **9,** 1281–1291 (2014).
18. Wolfson, R. L. *et al.* Sestrin2 is a leucine sensor for the mTORC1 pathway. *Science* **351,** 43–48 (2016).
19. Saxton, R. A. *et al.* Structural basis for leucine sensing by the Sestrin2-mTORC1 pathway. *Science* **351,** 53–58 (2016).
20. Kotaka, M., Ren, J., Lockyer, M., Hawkins, A. R. & Stammers, D. K. Structures of R- and T-state *Escherichia coli* aspartokinase III. Mechanisms of the allosteric transition and inhibition by lysine. *J. Biol. Chem.* **281,** 31544–31552 (2006).
21. Robin, A. Y. *et al.* A new mode of dimerization of allosteric enzymes with ACT domains revealed by the crystal structure of the aspartate kinase from Cyanobacteria. *J. Mol. Biol.* **399,** 283–293 (2010).
22. Dumas, R., Cobessi, D., Robin, A. Y., Ferrer, J.-L. & Curien, G. The many faces of aspartate kinases. *Arch. Biochem. Biophys.* **519,** 186–193 (2012).
23. Bridgham, J. T., Carroll, S. M. & Thornton, J. W. Evolution of hormone-receptor complexity by molecular exploitation. *Science* **312,** 97–101 (2006).
24. Coyle, S. M., Flores, J. & Lim, W. A. Exploitation of latent allostery enables the evolution of new modes of MAP kinase regulation. *Cell* **154,** 875–887 (2013).
25. Peisajovich, S. G., Garbarino, J. E., Wei, P. & Lim, W. A. Rapid diversification of cell signaling phenotypes by modular domain recombination. *Science* **328,** 368–372 (2010).
26. Zoncu, R., Efeyan, A. & Sabatini, D. M. mTOR: from growth signal integration to cancer, diabetes and ageing. *Nat. Rev. Mol. Cell Biol.* **12,** 21–35 (2011).
27. Shaw, R. J. & Cantley, L. C. Ras, PI(3)K and mTOR signalling controls tumour cell growth. *Nature* **441,** 424–430 (2006).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank all members of the Sabatini and Schwartz laboratories for helpful insights. This work is based on research conducted at the Northeastern Collaborative Access Team beamlines, which are funded by the National Institute of General Medical Sciences from the National Institutes of Health (P41 GM103403). The Pilatus 6M detector on 24-ID-C beam line is funded by a NIH-ORIP HEI grant (S10 RR029205). This research used resources of the Advanced Photon Source, a US Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under contract no. DE-AC02-06CH11357. This work has been supported by grants from NIH (R01CA103866 and AI47389) and the US Department of Defense (W81XWH-07- 0448) to D.M.S. Fellowship support was provided by NIH to L.C. (F31 CA180271). D.M.S. is an investigator of the Howard Hughes Medical Institute.

Author Contributions R.A.S., T.U.S., and D.M.S. designed the research plan. R.A.S. performed the experiments with assistance from L.C. and K.E.K. on experimental design and interpretation. R.A.S., T.U.S., and D.M.S. wrote the manuscript and all authors edited it.

Author Information Coordinates and structure factors for the x-ray crystal structure of CASTOR1 have been deposited in the Protein Data Bank (PDB) with accession code 5I2C. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of this paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.M.S. (sabatini@wi.mit.edu) or T.U.S. (tus@mit.edu).

Reviewer Information *Nature* thanks L. Tong and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Materials. Reagents were obtained from the following sources: HRP-labelled anti-rabbit secondary antibody from Santa Cruz Biotechnology; antibodies to phospho-T389 S6K1, S6K1, Mios and the FLAG epitope from Cell Signalling Technology; antibodies to the haemagglutinin epitope from Bethyl laboratories; antibody to raptor from Millipore. All antibodies used have been published previously[12,19]. FLAG–M2 affinity gel and amino acids from Sigma Aldrich; RPMI without leucine, arginine, or lysine from Pierce; DMEM from SAFC Biosciences; XtremeGene9 and Complete Protease Cocktail from Roche; inactivated fetal calf serum (IFS) from Invitrogen; [^{3}H]-labelled arginine from American Radiolabelled Chemicals.

Protein production and purification. Full-length, codon-optimized human CASTOR1 was N-terminally fused with a human rhinovirus 3C protease-cleavable His_{10}–Arg_8–ScSUMO tag and cloned into a pET-Duet-1 bacterial expression vector. This vector was transformed into *E. coli* LOBSTR (DE3) cells (Kerafast)[28]. Cells were grown at 37 °C to 0.6 optical density (OD), then protein production was induced with 0.2 mM IPTG at 18 °C for 12–14 h. Cells were collected by centrifugation at 6,000*g*, re-suspended in lysis buffer (50 mM potassium phosphate, pH 8.0, 500 mM NaCl, 30 mM imidazole, 3 mM β-mercaptoethanol (βME) and 1 mM PMSF) and lysed with a cell disruptor (Constant Systems). The lysate was cleared by centrifugation at 10,000*g* for 20 min. The soluble fraction was incubated with Ni-Sepharose 6 Fast Flow beads (GE Healthcare) for 30 min on ice. After washing of the beads with lysis buffer, the protein was eluted in 250 mM imidazole, pH 8.0, 150 mM NaCl and 3 mM βME. The Ni eluate was diluted 1:1 with 10 mM potassium phosphate, pH 8.0, 0.1 mM EDTA and 1 mM dithiothreitol (DTT), and was subjected to cation-exchange chromatography on a 5 ml SP sepharose fast flow column (GE Healthcare) with a linear NaCl gradient. The eluted CASTOR1 was then incubated with 3C protease and dialysed overnight at 4 °C into 10 mM potassium phosphate, pH 8.0, 150 mM NaCl, 0.1 mM EDTA and 1 mM DTT, followed by a second cation-exchange chromatography run on an SP Sepharose Fast Flow column (GE Healthcare) with a linear NaCl gradient. The protein was further purified via size-exclusion chromatography on a Superdex S200 16/60 column (GE Healthcare) equilibrated in running buffer (10 mM Tris-HCl, pH 8.0, 150 mM NaCl, 0.1 mM EDTA and 1 mM DTT). Selenomethionine (SeMet)-derivatized CASTOR1 was prepared as described previously[29] and purified as the native version, except that the reducing-agent concentration (βME and DTT) was 5 mM in all buffers.

Crystallization. Purified CASTOR1 was concentrated to 6 mg/ml and incubated in 2 mM arginine for >1 h before setting crystal trays. Crystals were grown at 18 °C by hanging-drop vapour diffusion with 1 µl of protein at 6 mg/ml mixed with an equal volume of reservoir solution containing 0.1 M sodium acetate pH 5.0, 0.25 M ammonium acetate, and 22.5% PEG 3350. Selenomethionine-derivatized CASTOR1 was crystallized in 0.1 M BIS-TRIS pH 5.6, 0.25 M ammonium acetate, and 22.5% PEG3350. Crystals were cryoprotected in mother liquor supplemented with 20% (v/v) ethylene glycol.

Data collection and structure determination. Data collection was performed at the Advanced Photon Source end station 24-IDC at Argonne National Laboratory, at 100 K. All data-processing steps were carried out with programs provided through SBgrid[30]. Data reduction was performed with HKL2000[31]. A complete native data set was collected to 1.8 Å (at wavelength 0.9792 Å) and a complete SeMet data set, at the selenium peak wavelength (0.9792 Å), was collected to 2.2 Å. The phase problem was solved using single-wavelength anomalous dispersion (SAD) and selenium positions were determined in HySS, run as part of the PHENIX AutoSol program[32], for the SeMet data set (space group $P2_1$, 4 molecules per asymmetric unit). An interpretable 2.2 Å experimental electron density map was obtained, and manual model building was carried out in Coot[33]. Subsequent refinement was carried out with the superior 1.8 Å native data set using *phenix.refine* to a final R_{work}/R_{free} of 17.2%/20.4%. Ramachandran statistics in the final model are 99% favoured, 1% allowed, and 0% outlier.

Structural analysis. Protein–protein and protein–ligand interfaces were analysed using PDBePISA[34]. NCBI's Vector Alignment Search Tool (VAST)[35] was used to identify structurally related proteins in the PDB. The multiple sequence alignment (MSA) was generated in Jalview[36] with the T-Coffee alignment algorithm[37]. Sequences of CASTOR1 homologues were obtained via NCBI BLAST searches[38]. All structure figures were made in PyMol[39].

Cell lysis and immunoprecipitation. Cells were rinsed once with ice-cold PBS and immediately lysed with Triton lysis buffer (1% Triton, 10 mM β-glycerol phosphate, 10 mM pyrophosphate, 40 mM HEPES pH 7.4, 2.5 mM $MgCl_2$ and 1 tablet of EDTA-free protease inhibitor (Roche) (per 25 ml buffer). The cell lysates were cleared by centrifugation at 13,000 rpm at 4 °C in a microcentrifuge for 10 min. For anti-HA-immunoprecipitations, the magnetic anti-HA beads (Pierce) were washed three times with lysis buffer. 30 µl of a 50/50 slurry of the affinity gel was then added to clarified cell lysates and incubated with rotation for 1 h at 4 °C. Following immunoprecipitation, the beads were washed four times with lysis buffer containing 500 mM NaCl. Immunoprecipitated proteins were denatured by the addition of 50 µl of sample buffer and boiling for 5 min as described[40], resolved by 8–16% SDS–PAGE, and analysed by immunoblotting.

For co-transfection experiments in HEK-293T cells, 2.5 million cells were plated in 10 cm culture dishes. Twenty-four hours later, cells were transfected using the polyethylenimine method[41] with the pRK5-based cDNA expression plasmids indicated in the following amounts: 50 ng CASTOR1–HA (wild-type or mutant), 50 ng CASTOR1–FLAG, 1 µg HA–metap2, or 2 ng S6K. For *in vitro* dissociation experiments, 50 ng of wild-type CASTOR1–HA was transfected into HEK-293T cells. The total amount of plasmid DNA in each transfection was normalized to 5 µg with empty pRK5. 36–48 h after transfection, cells were lysed as described above.

For experiments that required amino acid starvation or re-stimulation, cells were treated as previously described[42]. Briefly, cells were incubated in arginine-free RPMI for 50 min and then re-stimulated with 500 µM arginine for 10 min.

Arginine binding assay. Five million HEK-293T cells were plated on a 15 cm plate four days before the experiment. Twenty-four hours after plating, the cells were transfected via the polyethylenimine method with the pRK5-based cDNA expression plasmids indicated in the figures in the following amounts: 15 µg FLAG–Rap2A, 500 ng FLAG–CASTOR1 (wild-type or mutant). The total amount of plasmid DNA in each transfection was normalized to 15 µg total DNA with empty PRK5. Forty-eight hours after transfection cells were lysed as previously described. If multiple samples of the same type were represented in the experiment, the cell lysates were combined, mixed, and evenly distributed amongst the relevant tubes.

Anti-FLAG beads were blocked by rotating in 1 µg/µl bovine serum albumin (BSA) for 20 min at 4 °C, then washed twice in lysis buffer and re-suspended in an equal volume of lysis buffer. 30 µl of bead slurry was added to each of the clarified cell lysates and incubated as previously described. After immunoprecipitation, the beads were washed as previously and incubated for one hour on ice in cytosolic buffer (0.1% Triton, 40 mM HEPES pH 7.4, 10 mM NaCl, 150 mM KCl, 2.5 mM $MgCl_2$) with the appropriate amount of [^{3}H]-labelled arginine and cold arginine. At the end of one hour, the beads were aspirated dry and rapidly washed three times with cytosolic buffer. The beads were aspirated dry again and resuspended in 85 µl of cytosolic buffer. Each sample was mixed well and three 10 µl aliquots were quantified separately using a TriCarb scintillation counter (PerkinElmer). This process was repeated in pairs for each sample, to ensure similar incubation and wash times for all samples analysed across different experiments.

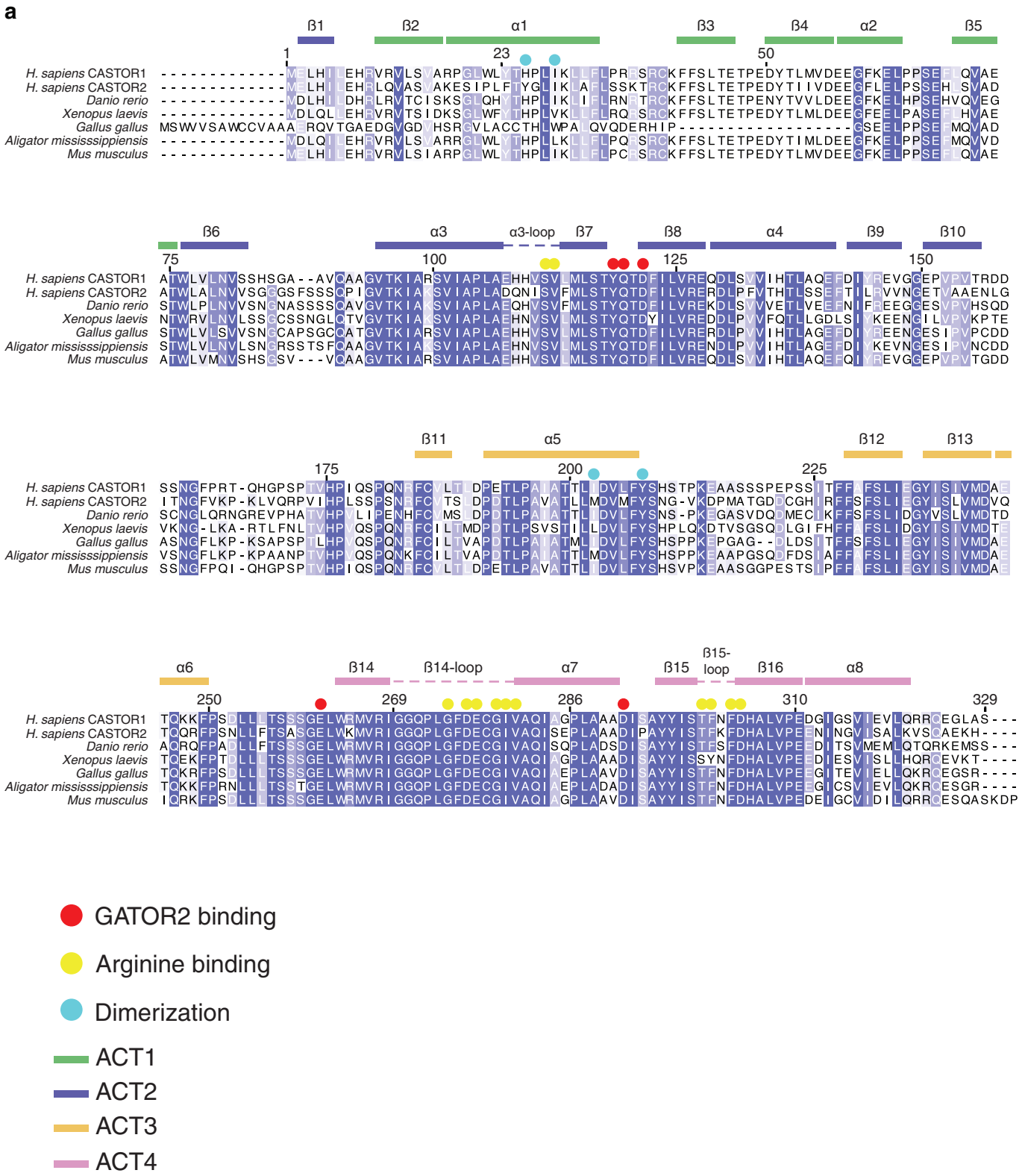
***In vitro* CASTOR1–GATOR2 dissociation assay with arginine analogues.** HEK-293T were transfected with HA–CASTOR1 constructs as described above. 48 h after transfection, cells were starved for all amino acids for 50 min, lysed and subjected to anti-FLAG immunoprecipitation as described previously. The CASTOR1–GATOR2 complexes immobilized on the haemagglutinin beads were washed twice in lysis buffer with 500 mM NaCl, then incubated for 20 min in 1 ml of cytosolic buffer with 400 µM of the indicated compound. The amount of GATOR2 and CASTOR1 that remained bound was assayed by SDS–PAGE and immunoblotting as described previously.

Cell lines and tissue culture. HEK-293T cells were maintained at 37 °C and 5% CO2 and cultured in DMEM 10% IFS supplemented with 2 mM glutamine, penicillin (100 IU/ml) and streptomycin (100 µg/ml). HEK-293T cells were obtained from the American Type Culture Collection (ATCC) and were free of mycoplasma contamination.

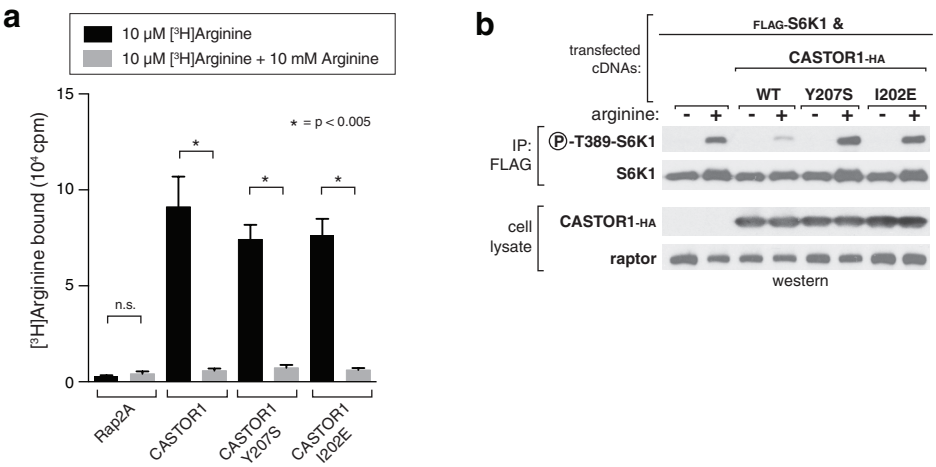
Statistical analysis. For the arginine-binding assays, two-tailed *t*-tests were used for comparison between two groups. All comparisons were two-sided, and *P* values of less than 0.005 were considered statistically significant. The data meet the assumptions of the test and the variance is similar between groups that are being statistically compared. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

28. Andersen, K. R., Leksa, N. C. & Schwartz, T. U. Optimized E. coli expression strain LOBSTR eliminates common contaminants from His-tag purification. *Proteins* **81,** 1857–1861 (2013).
29. Brohawn, S. G., Leksa, N. C., Spear, E. D., Rajashankar, K. R. & Schwartz, T. U. Structural evidence for common ancestry of the nuclear pore complex and vesicle coats. *Science* **322,** 1369–1373 (2008).
30. Morin, A. *et al.* Collaboration gets the most out of software. *eLife* **2,** e01456 (2013).
31. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
32. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66,** 213–221 (2010).
33. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66,** 486–501 (2010).

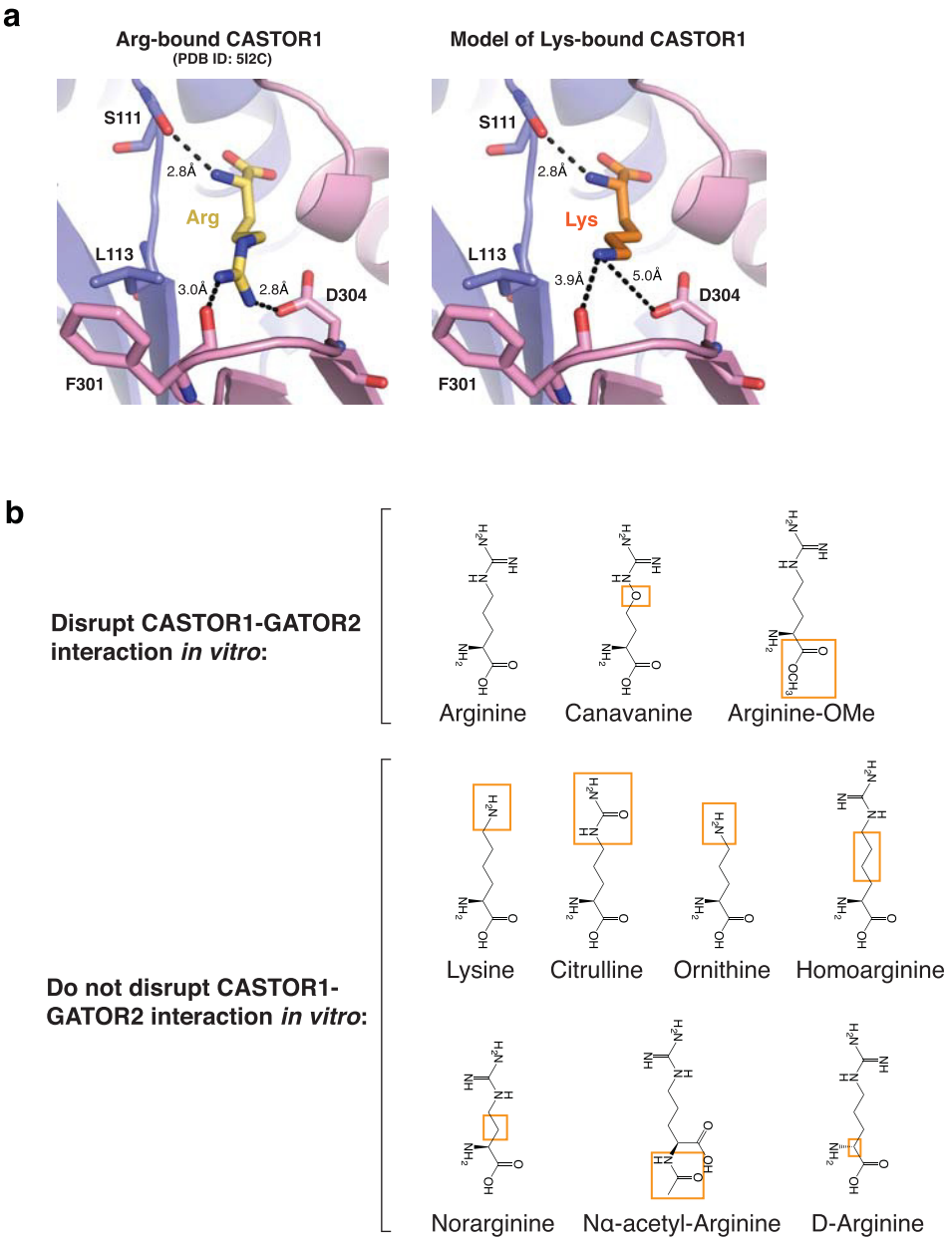
34. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372,** 774–797 (2007).
35. Gibrat, J. F., Madej, T. & Bryant, S. H. Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **6,** 377–385 (1996).
36. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25,** 1189–1191 (2009).
37. Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302,** 205–217 (2000).
38. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
39. Schrodinger, LLC. The PyMOL Molecular Graphics System, Version 1.3r1 (2010).
40. Kim, D.-H. *et al.* mTOR interacts with raptor to form a nutrient-sensitive complex that signals to the cell growth machinery. *Cell* **110,** 163–175 (2002).
41. Boussif, O. *et al.* A versatile vector for gene and oligonucleotide transfer into cells in culture and in vivo: polyethylenimine. *Proc. Natl Acad. Sci. USA* **92,** 7297–7301 (1995).
42. Tsun, Z.-Y. *et al.* The folliculin tumor suppressor is a GAP for the RagC/D GTPases that signal amino acid levels to mTORC1. *Mol. Cell* **52,** 495–505 (2013).



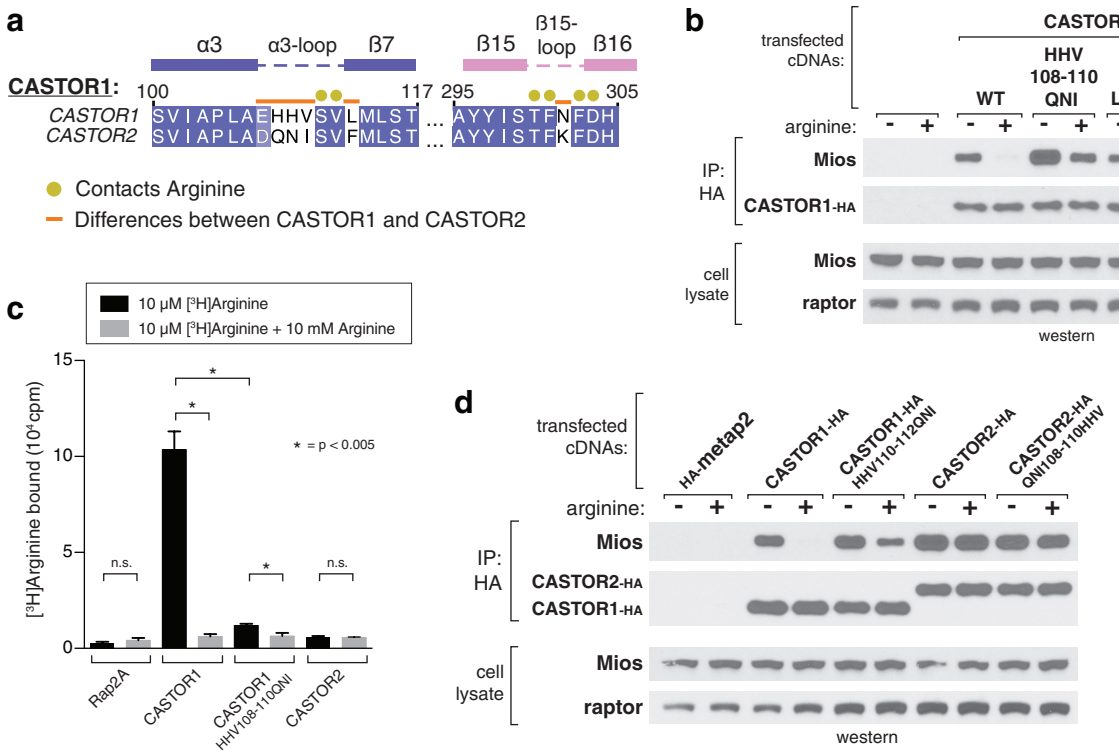
Extended Data Figure 1 | Multiple sequence alignment of CASTOR1 homologues. a, Expanded Multiple Sequence Alignment of CASTOR1 homologues from various organisms. Positions are coloured white to blue according to increasing sequence identity. Secondary structure features are labelled and coloured by ACT domain as in Fig. 1a.



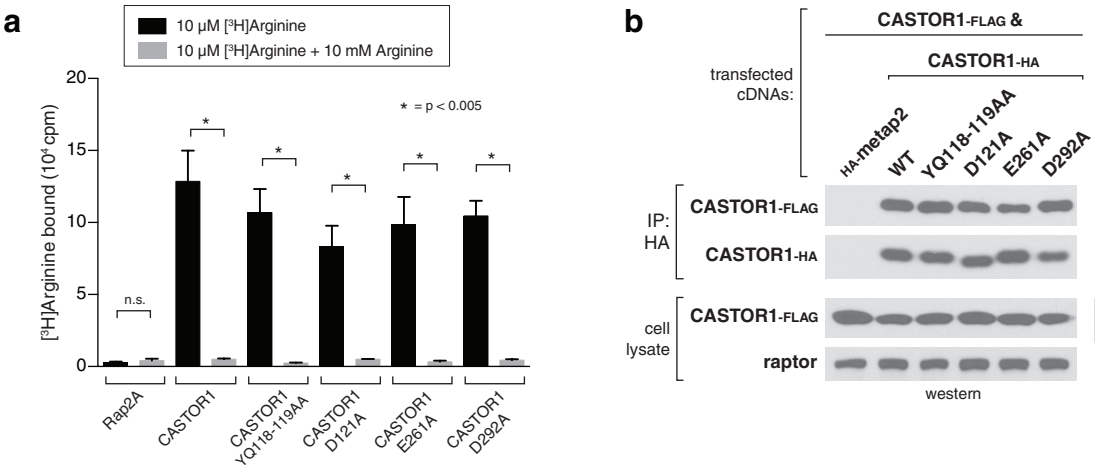
Extended Data Figure 2 | Dimerization-deficient CASTOR1 mutants bind arginine but fail to inhibit mTORC1 in cells. a, The dimerization-deficient CASTOR1 Y207S and I202E mutants bind arginine *in vitro*. FLAG-immunoprecipitates prepared from HEK-293T cells transiently expressing indicated FLAG-tagged proteins were used in binding assays with [^{3}H]Arginine as described in the Methods. Unlabelled arginine was included as a competitor where indicated. Values are mean ± s.d. for three technical replicates from one representative experiment. **b**, Dimerization-deficient CASTOR1 Y207S and I202E mutants fail to inhibit mTORC1. HEK-293T cells transiently expressing FLAG–S6K1 and HA-tagged wild-type, Y207S, or I202E CASTOR1 were starved of arginine for 50 min and, where indicated, re-stimulated for 10 min. FLAG-immunoprecipitates were prepared from lysates and analysed as in Fig. 1c. Phospho-S6K1 was used as an indicator of mTORC1 activity.



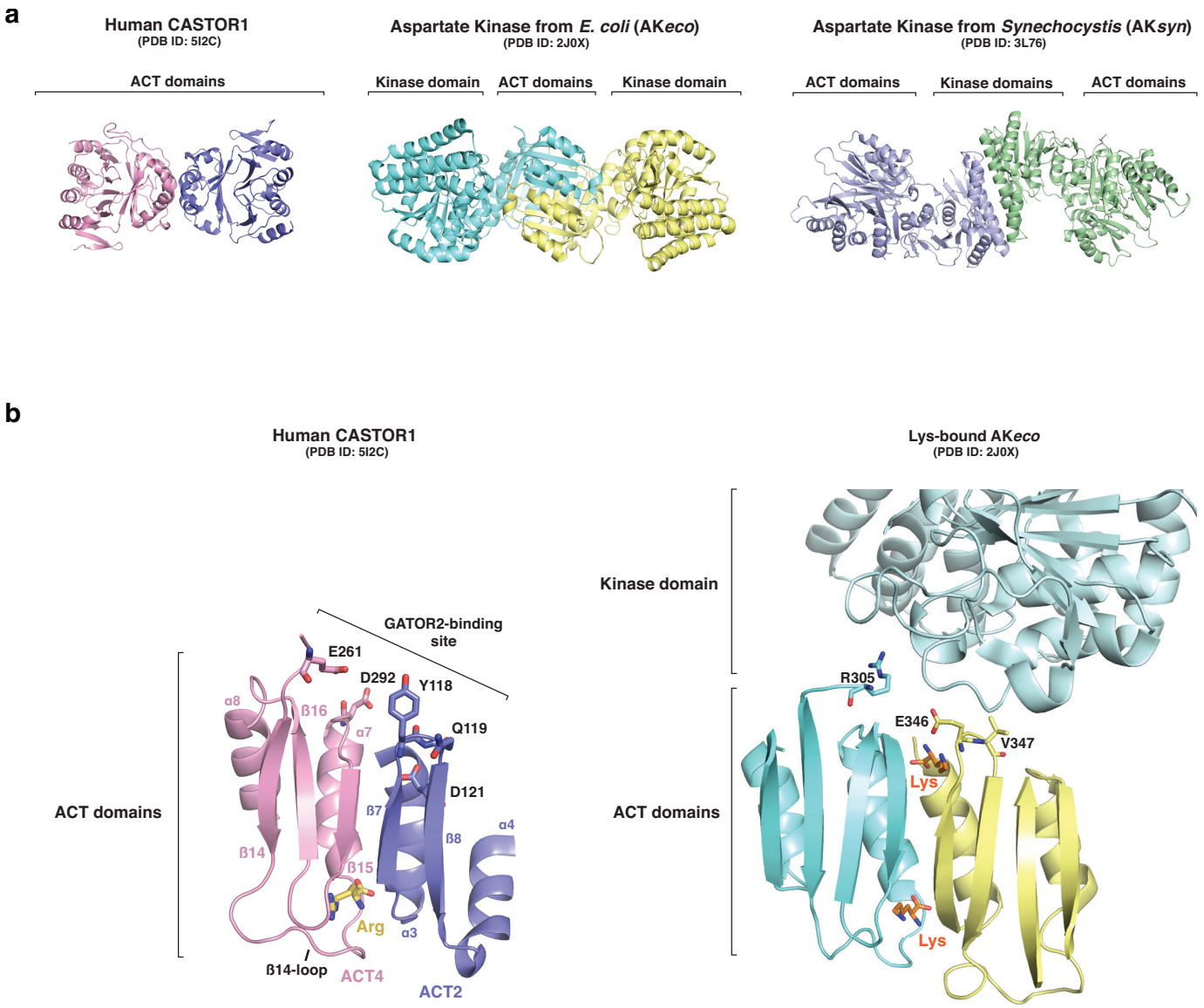
Extended Data Figure 3 | Model of lysine-binding in CASTOR1. **a**, Comparison of the arginine-bound pocket of human CASTOR1 with a model of the pocket with lysine in place of arginine. Arginine and lysine stick representations are shown in yellow and orange, respectively. The distances in the lysine-bound model, 3.8 Å and 5.0 Å, are beyond the range of standard hydrogen bonds and salt bridges, respectively. ACT domains are labelled as in Fig. 1a. **b**, Chemical structures of arginine analogues used in Fig. 2e. Differences relative to L-arginine are highlighted in oranges boxes.



Extended Data Figure 4 | Differences in the arginine-binding capacities of CASTOR1 and CASTOR2. a, Multiple sequence alignment of human CASTOR1 and CASTOR2, highlighting differences in amino acid sequence that are in close proximity to arginine-binding residues in CASTOR1. **b**, The CASTOR1 HHV108–110QNI mutant constitutively binds GATOR2 in cells. HEK-293T cells transiently expressing HA–metap2 or the indicated HA-tagged CASTOR1 constructs were starved of arginine for 50 min and, where indicated, re-stimulated for 10 min. HA-immunoprecipitates were prepared and analysed as in Fig. 1c. **c**, The CASTOR1 HHV108–110QNI mutant displays reduced arginine-binding capacity *in vitro*. Binding assays were performed with the indicated CASTOR1 or CASTOR2 constructs and immunoprecipitates analysed as in Fig. 2c. Values are mean ± s.d. for three technical replicates from one representative experiment. **d**, Comparison of the CASTOR1 HHV108–110QNI mutant and wild-type CASTOR2. HEK-293T cells transiently expressing HA–metap2 or the indicated HA-tagged CASTOR1 or CASTOR2 constructs were starved of arginine for 50 min and, where indicated, re-stimulated for 10 min. HA-immunoprecipitates were prepared and analysed as in Fig. 1c.



Extended Data Figure 5 | GATOR2-binding-deficient CASTOR1 mutants still bind arginine and homodimerize. a, The CASTOR1 YQ118–119AA, D121A, E261A and D292A mutants bind arginine *in vitro*. FLAG-immunoprecipitates prepared from HEK-293T cells transiently expressing indicated FLAG-tagged proteins were used in binding assays with [^{3}H]arginine as described in the Methods. Unlabelled arginine was included as a competitor where indicated. Values are mean ± s.d. for three technical replicates from one representative experiment. **b**, The CASTOR1 YQ118–119AA, D121A, E261A and D292A mutants dimerize in cells. HA-immunoprecipitates prepared from HEK293T-cells transiently expressing CASTOR1–FLAG and HA–metap2 or the indicated HA-tagged CASTOR1 constructs were analysed as in Fig. 1c.



Extended Data Figure 6 | Similarities between human CASTOR1 and prokaryotic aspartate kinases. a, Ribbon diagram views of human CASTOR1, AKeco (PDB ID: 2J0x) and AKsyn (PDB ID: 3L76), highlighting the different modes of dimerization. Aspartate kinases can dimerize through an interlocked-ACT domain conformation (as in AKeco) or through their kinase domains (AKsyn), both of which are distinct from the side-by-side ACT-domain dimerization in CASTOR1. **b**, View of AKeco depicting positions of residues R305, E346, and V347, which correspond to the positions of the GATOR2-interacting residues of CASTOR1.

Extended Data Table 1 | Data collection and refinement statistics (SAD)

	CASTOR1 + Arg Native	CASTOR1 + Arg SeMet
Organism	*H. sapiens*	*H. sapiens*
PDB ID	5I2C	
Data collection		
Space group	$P2_1$	$P2_1$
Cell dimensions		
a, *b*, *c* (Å)	91.39, 82.60, 96.67	91.76, 82.35, 96.71
α, β, γ (°)	90, 116.23, 90	90, 116.04, 90
		Peak
Wavelength (Å)	0.9792	0.9792
Resolution (Å)	86.7 – 1.80	86.89 – 2.20
R_{sym} (%)	7.2 (62.8)	10.4 (>100)
$I / \sigma I$	25.9 (1.2)	22.6 (1.4)
Completeness (%)	97.85 (87.1)	98.2 (98.1)
Redundancy	3 (2.5)	6.4 (5.9)
Anomolous Completeness (%)		96.8
Refinement		
Resolution (Å)	86.71 – 1.80	
No. reflections	116,883	
R_{work} / R_{free}	17.2%/20.4%	
No. atoms	9,872	
Protein	9,012	
Arg	48	
Water	796	
Average *B*-factors (Å)	40.2	
Protein	40.0	
Arg	26.8	
Water	46.4	
R.m.s. deviations		
Bond lengths (Å)	0.007	
Bond angles (°)	0.85	

*Values in parentheses are for highest-resolution shell.

LETTER

doi:10.1038/nature19080

Reconstruction of bacterial transcription-coupled repair at single-molecule resolution

Jun Fan[1], Mathieu Leroux-Coyau[1], Nigel J. Savery[2] & Terence R. Strick[1,3,4]

Escherichia coli **Mfd translocase enables transcription-coupled repair by displacing RNA polymerase (RNAP) stalled on a DNA lesion and then coordinating assembly of the UvrAB(C) components at the damage site[1–4]. Recent studies have shown that after binding to and dislodging stalled RNAP, Mfd remains on the DNA in the form of a stable, slowly translocating complex with evicted RNAP attached[5,6]. Here we find, using a series of single-molecule assays, that recruitment of UvrA and UvrAB to Mfd–RNAP arrests the translocating complex and causes its dissolution. Correlative single-molecule nanomanipulation and fluorescence measurements show that dissolution of the complex leads to loss of both RNAP and Mfd. Subsequent DNA incision by UvrC is faster than when only UvrAB(C) are available, in part because UvrAB binds 20–200 times more strongly to Mfd–RNAP than to DNA damage. These observations provide a quantitative framework for comparing complementary DNA repair pathways *in vivo*.**

The conformational changes that take place in Mfd upon docking to, and activation by, stalled RNAP[7–9] enable it to bind to DNA upstream of RNAP and translocate along DNA against stalled RNAP[10,11], and to expose a UvrB homology module and recruit UvrA[3]. Remarkably, single-molecule assays have shown that after displacing stalled RNAP to make the lesion accessible for repair, Mfd continues to translocate slowly and processively with RNAP attached to it[5,6]. These assays help explain recent results showing that transcription-coupled repair (TCR) can also accelerate repair of damaged sites downstream of the stall site[12]. Nevertheless, the role of the translocating Mfd–RNAP complex in stimulating repair by UvrAB(C) remains unclear. Here three single-molecule assays based on magnetic trapping[13] are brought to bear on the system.

In the tethered-RNAP translocation assay we stalled biotinylated RNAP after transcribing 20 bases using only ATP, UTP and GTP on a DNA cassette lacking cytosine residues. We then tethered the RNAP to a streptavidin-coated magnetic bead, and anchored the linear DNA template at one end to a modified glass coverslip. We thus obtained an RNAP stalled ~1 kilobase pair (kbp) from one end of an ~8 kbp DNA as it transcribed towards the distant glass surface[14] (Fig. 1a). The DNA was extended away from the surface by a vertical force ($F = 1$ pN) applied to the bead using a pair of magnets located above the sample, and the bead's position above the surface was detected in real time using computer-aided videomicroscopy. Addition of 100 nM Mfd and 2 mM ATP caused motion of the bead towards the surface as an Mfd–RNAP complex formed and translocated along the DNA (Fig. 1b)[5,6]. The complex was Michaelian with respect to ATP, with maximum rate $V_{max}^{ATP} = 4.7 \pm 0.1$ bp s^{-1} (s.e.m.) and

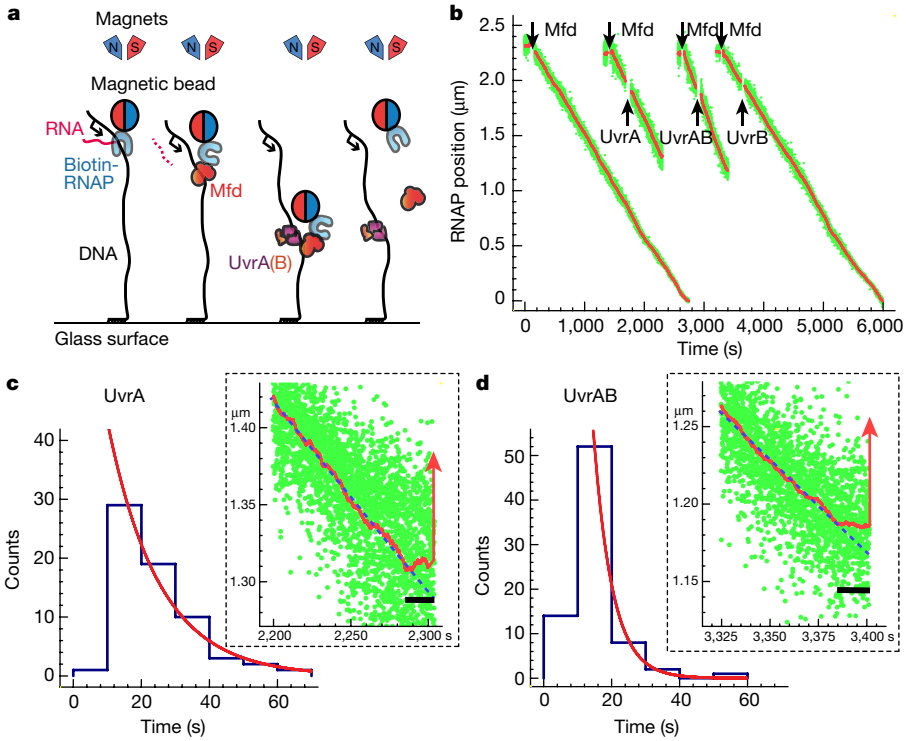


Figure 1 | Tethered-RNAP assay for resolution of the Mfd–RNAP complex. a, UvrA(B) intercepts, arrests and releases translocating Mfd–RNAP-bead complexes (see text). **b**, Time-traces of bead position in the presence of 2 mM ATP and proteins as indicated. Arrows indicate component infusion (gaps). **c**, Single-exponential lifetime distribution of arrest events in the presence of UvrA displays a mean of 15 ± 3 s (s.e.m., $n = 65$). Inset time-trace shows Mfd–RNAP arrest (black bar) and release (red up-arrow) for 10 s averaging. Dashed line, linear fit to translocation. **d**, As in **c** but in the presence of UvrA and UvrB. Mean arrest time now 6 ± 1 s (s.e.m., $n = 77$).

[1]Institut Jacques Monod, CNRS, UMR7592, University Paris Diderot, Sorbonne Paris Cité F-75205 Paris, France. [2]DNA-Protein Interactions Unit, School of Biochemistry, University of Bristol, Bristol BS8 1TD, UK. [3]Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS, Inserm, Ecole Normale Supérieure, PSL Research University, F-75005 Paris, France. [4]Programme Equipe Labellisées, Ligue Contre le Cancer, 75013 Paris, France.

Table 1 | Statistics of release of the Mfd–RNAP complex in the translocation assay

	Release before surface		Dissociation by surface collision	Efficiency of release (%)
	Pausing detected	No pausing detected		
Mfd–RNAP alone	0	33	66	33
+UvrA	65	67	43	75
+UvrAB	63	208	20	93
+UvrB	0	5	65	7

Michaelis constant $K_m^{ATP} = 16 \pm 0.4\,\mu M$ (s.e.m.) (Extended Data Fig. 1). In ~50% of cases the bead translocated ~7,000 bp under the action of Mfd and was released only upon collision with the surface; in the remaining cases it released before reaching the surface (Table 1).

Addition of 50 pM UvrA to translocating Mfd–RNAP (Fig. 1b) led to release of 75% of beads before reaching the surface (Table 1). In 50% of release events we observed arrest of the translocating complex before release (65/132; Table 1). Arrest duration was well described by single-exponential kinetics with a mean duration of 15 ± 3 s (s.e.m.; Fig. 1c). Accordingly, arrest most probably occurred in the remaining 50% of events but was too short for us to detect with a slow measurement response time of ~10 s (see Supplementary Information: Methods). Addition of 50 pM UvrA and 250 nM UvrB to the translocating complex led to release of 93% of beads (Fig. 1b and Table 1). Arrest now lasted on average only 6 ± 1 s (s.e.m.; Fig. 1d), and was observed in only ~20% of cases (Table 1). As above, shorter periods of arrest were predicted to have taken place in all cases. UvrB alone failed to destabilize the translocating complex (Fig. 1b and Table 1). ATP hydrolysis was required for either UvrA or UvrAB to perform these tasks (Extended Data Fig. 2). These results indicate that complex resolution by UvrAB is faster and more efficient than by UvrA alone, and that arrest of the complex is on-pathway to its disassembly.

We next used a tethered-DNA assay in which a 2 kbp subfragment of the DNA used previously was attached to a magnetic bead and a coverslip, and extended ($F = 0.3$ pN) and supercoiled in the magnetic trap (Fig. 2a)[13,15]. Fig. 2b shows the extension signal obtained when RNAP initiated transcription and stalled on a positively supercoiled DNA substrate bearing a cyclobutane pyrimidine dimer (CPD), and was displaced by Mfd to form the long-lived translocating complex (denoted intermediate, I; Fig. 2b, Extended Data Fig. 3 and ref. 5). Intermediate lifetimes were long and normally distributed (mean of 548 ± 37 s s.e.m.), and independent of supercoiling or cause of stalling (Extended Data Fig. 3), but depended on the distance between stalled RNAP and the end of the DNA (compare Extended Data Figs. 3 and 4). Addition of 50 pM UvrA and 250 nM UvrB reduced the mean lifetime of the Mfd–RNAP repair intermediate (Fig. 2c) to 141 ± 20 s (s.e.m.) in a manner that was essentially unaffected by supercoiling and cause of stalling (Extended Data Figs 4 and 5). The lifetime distribution of the intermediate species then followed a difference-of-exponentials function characteristic of a Michaelis–Menten process with association/dissociation of UvrAB to translocating Mfd–RNAP (rates k_1 and k_{-1} respectively) and a slow forward catalytic rate for resolution of the complex k_2.

By titrating UvrA from 50 to 100 pM against a saturating concentration of 250 nM UvrB, we observed a gradual reduction in the mean lifetime of the Mfd–RNAP–DNA complex as expected for a diffusion-limited process (see Fig. 2d and Extended Data Fig. 6). By determining the mean lifetimes of intermediates for different UvrA concentrations and fitting those average values to a Michaelis–Menten model, we obtained $K_m^{UvrAB} = 96 \pm 51$ pM (s.e.m.) and $V_{max}^{UvrAB} = 0.023 \pm 0.006\,s^{-1}$ ($1/V_{max}^{UvrAB} = 43 \pm 12$ s (s.e.m.)) for dissolution of the Mfd–RNAP complex by UvrAB (Fig. 2d). By globally fitting the full lifetime distributions of Mfd–RNAP intermediates obtained at different UvrA concentrations to the single-molecule limit for the Michaelis–Menten equation[16], using the maximum velocity obtained above as a global constraint (Extended Data Fig. 6), we estimated the on- and off-rates of the UvrAB complex with respect to the Mfd–RNAP intermediate as $k_1 = 7.3 \pm 1.9 \times 10^8\,M^{-1}\,s^{-1}$, and $k_{-1} = 0.037 \pm 0.013\,s^{-1}$, giving a value of K_m^{UvrAB} of about 80 pM, in agreement with the estimate from average values. Control experiments showed that removing Mfd, ATP or UvrA abolishes complex dissolution (Extended Data Fig. 7). UvrB could not be removed, however, as UvrA, on its own, compacted DNA in a non-specific manner at concentrations as low as 10 pM, precluding its analysis unbuffered by UvrB in this assay (Extended Data Fig. 8). Importantly, DNA compaction by 100 pM UvrA was abolished by addition of 250 nM UvrB (Extended Data Fig. 8).

To further determine the fate of the Mfd–RNAP complex upon arrest and dissolution by UvrAB we used NanoCOSM[6], a recently developed assay enabling correlative nanomanipulation and fluorescence co-localization of single molecules. By tracking the Mfd–RNAP intermediate via its mechanical signature on DNA as seen in the topological assay, and simultaneously using single-molecule fluorescence to identify the co-localization of labelled components, we could monitor the composition of the repair complex as it progressed through TCR. Fluorescently labelled RNAP appeared in the fluorescence channel when transcription initiation was observed

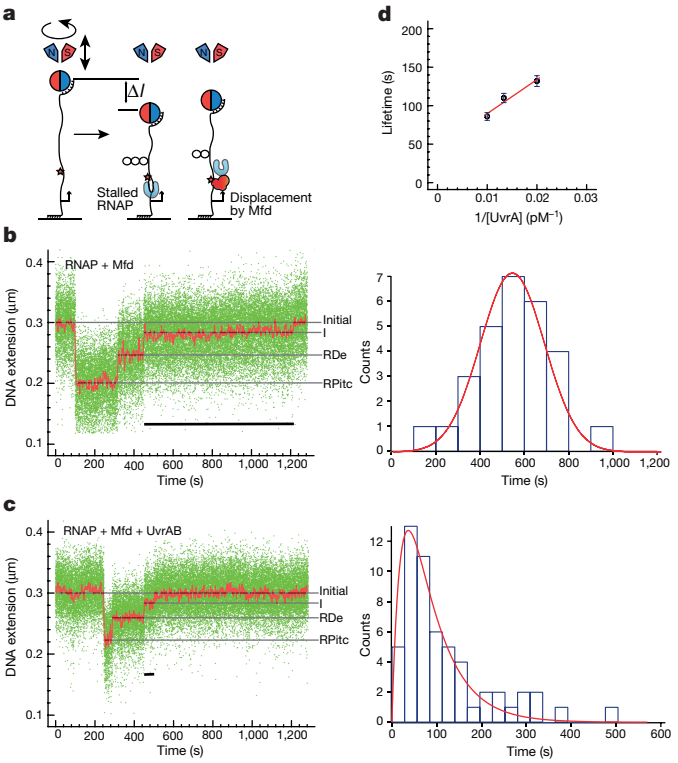


Figure 2 | Tethered-DNA assay for resolution of the Mfd–RNAP complex. **a**, Transcription complexes can be monitored as the positively supercoiled DNA couples local torsional deformation by RNAP into large-scale looping (writhe) deformation[21]. **b**, Left: tethered-DNA time-trace from CPD-bearing DNA in the presence of RNAP, Mfd, GreB, UTP, GTP, CTP and 2 mM ATP shows formation of initially transcribing complex (RPitc), stalled elongation complex (RDe) and Mfd–RNAP repair intermediate (I, underscored by black bar) which resolves to baseline. Right: lifetime of intermediate follows a Gaussian distribution (red line; $n = 28$). **c**, As in **b** but adding UvrA and UvrB ($n = 58$). Right: red line is the predicted distribution based on kinetic constants (main text). **d**, Intermediate lifetime plotted as a function of inverse concentration of UvrA, obtained as in **c** but for RNAP stalled on a positively supercoiled cytosine (C)-less cassette.

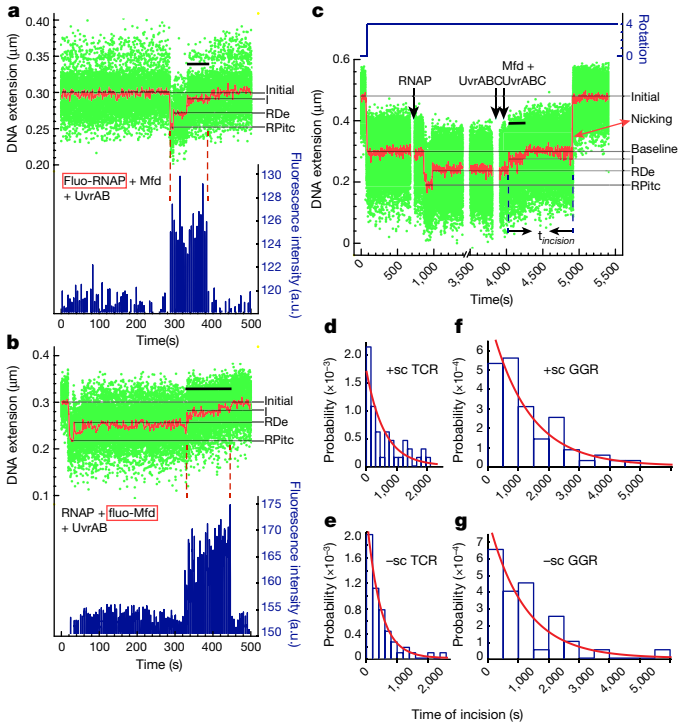


Figure 3 | Correlative single-molecule analysis of Mfd–RNAP handoff to UvrAB, and downstream DNA incision by UvrC. Time-traces showing simultaneous tethered-DNA assay of Mfd–RNAP repair intermediate formation and resolution by UvrAB, and single-molecule fluorescence signals from experiments in which (**a**) RNAP ($n = 14$) or (**b**) Mfd ($n = 21$) were fluorescently labelled (Fluo-RNAP or Fluo-Mfd[6]). The traces presented were obtained on C-less cassette DNA. **c**, Tethered-DNA time-trace for the TCR incision assay performed on positively supercoiled, CPD-bearing DNA. Down-arrows: stalled RNAP at lesion; pre-equilibrate with UvrA, UvrB, UvrC, pUC18 DNA, GreB and nucleoside 5′-triphosphates (NTPs); add 100 nM Mfd and 1 mM ATP (see text and Supplementary Information). Incision time distribution and exponential fits for (**d**, **e**) TCR and (**f**, **g**) GGR on positively or negatively supercoiled CPD-containing DNA, respectively.

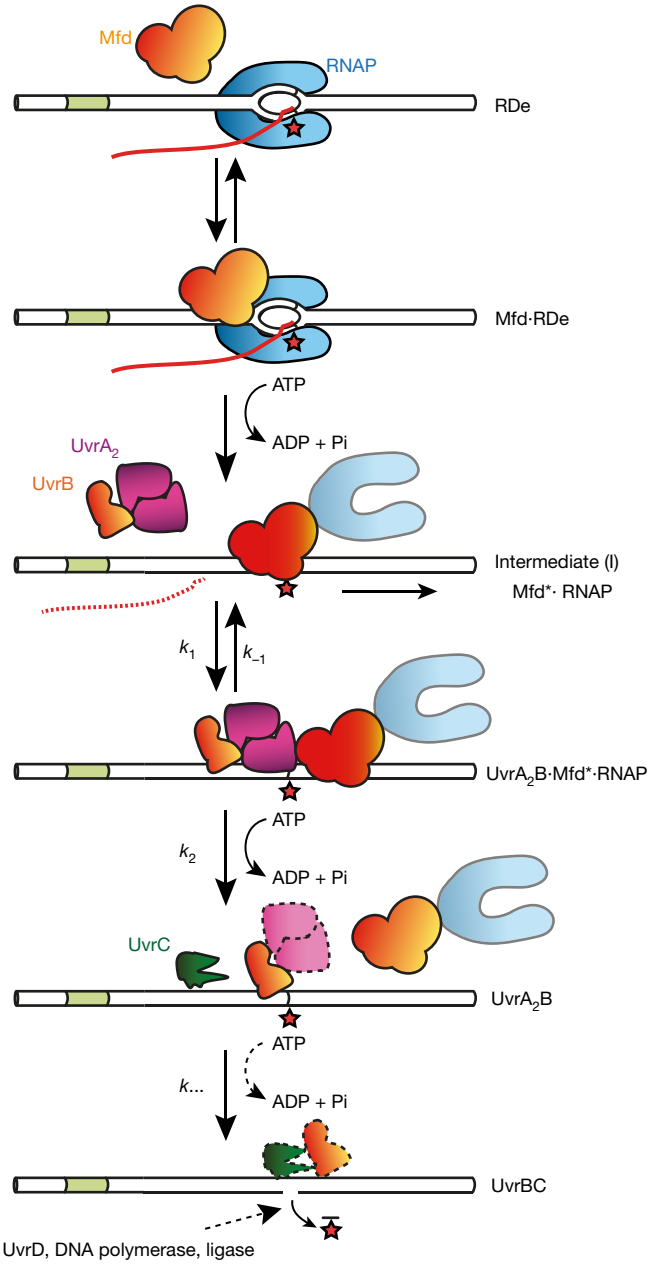


Figure 4 | A model for the TCR pathway. Stalled RNAP recruits and activates Mfd (denoted Mfd*), which then displaces RNAP from the lesion. Upon displacement of stalled RNAP, Mfd* remains attached to RNAP and continues to slowly translocate along the DNA. Mfd* can in turn recruit UvrAB, leading to arrest and release of the Mfd–RNAP complex from DNA in an ATP-dependent manner. We propose that this places UvrAB in the immediate vicinity of the lesion, ultimately increasing the rate at which UvrAB finds the lesion. UvrB-DNA can then recruit UvrC, which incises the lesion site.

nanomechanically and was lost from that channel upon nanomechanical dissolution by UvrAB of the repair intermediate (Fig. 3a and Extended Data Fig. 9a). Similarly, fluorescently labelled Mfd appeared in the fluorescence channel upon formation of the repair intermediate, and was lost from that channel upon dissolution by UvrAB of the repair intermediate (Fig. 3b and Extended Data Fig. 9b). Thus dissolution of the stable Mfd–RNAP intermediate by UvrAB involves loss of both Mfd and RNAP, indicating they do not act in downstream steps of DNA repair.

We finally used the tethered-DNA assay to measure TCR incision rates. DNA incision resulted in an abrupt loss of supercoiling and was readily detectable as a sudden increase in end-to-end extension. TCR incision was obtained by first stalling RNAP on a CPD, then pre-equilibrating the cell with UvrAB(C), and finally rapidly adding Mfd to the system (see Fig. 3c). We measured $t_{incision}$, the time elapsed between remodelling of stalled RNAP by Mfd and DNA incision by UvrC. Incision times for positively and negatively supercoiled DNA followed single-exponential distributions with mean lifetimes of 380 ± 120 s (s.e.m., $n = 44$ events) and 390 ± 70 s (s.e.m., $n = 59$ events), respectively (Fig. 3d, e). This is significantly faster than incision rates of the CPD substrate in the presence of only UvrABC, as in the case of global genome repair (GGR) (1,230 ± 195 s (s.e.m.), $n = 72$, and 1,156 ± 256 s (s.e.m.), $n = 40$, for positive and negative supercoiling, respectively; see Fig. 3f, g and Extended Data Fig. 10 for experiments and controls removing ATP, UvrAB, the CPD or in which CPD is protected by RNAP). Our observation of an enhanced repair rate in this assay, even after accounting for GGR rates, is consistent with previous biochemical findings[12,17].

TCR differs from GGR in the mode of recruitment of UvrAB to the lesion. UvrAB on its own displays reasonable affinity for DNA damage (dissociation constant in the 1–10 nM range[18]). Here we have shown that the dissociation constant of UvrAB from the activated Mfd–RNAP complex is 20–200 times smaller ($k_{-1}/k_1 = 50$ pM). This is partly due to the extremely efficient docking of UvrAB to the exposed UvrB homology module of the Mfd–RNAP complex ($k_1 \approx 7 \times 10^8$ M s^{-1}, essentially diffusion-limited). This efficiency can be explained by the fact that the UvrB homology module is larger and more accessible and thus

easier to 'find' than DNA damage. In this manner, GGR components may actively participate in repair even in uninduced, non-SOS conditions, where abundance of UvrA is extremely low—in the 20 nM range: that is, only about ten dimers per cell[19,20]—and therefore also subject to large fluctuations. The TCR pathway we detail therefore appears to be most relevant to 'housekeeping' DNA repair, watchfully maintaining genomic integrity even in the absence of stressful or genotoxic conditions.

Our observations suggest the existence of a transient UvrB–UvrA–UvrA–Mfd–RNAP repair complex, which would convert into a UvrB–UvrA–UvrA complex after loss of Mfd and RNAP (see model in Fig. 4). Intriguingly, this complex is able to drive repair only of the transcribed strand of DNA—the hallmark of TCR[12]. This suggests the complex, once loaded onto the DNA, does not 'pick up' a second UvrB. The single-molecule methods used here provide us not only with both broad and detailed views of TCR, but also with the opportunity to pursue these advanced mechanistic questions.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 16 February; accepted 5 July 2016.
Published online 3 August 2016.

1. Witkin, E. M. Radiation-induced mutations and their repair. *Science* **152,** 1345–1353 (1966).
2. Mellon, I., Spivak, G. & Hanawalt, P. C. Selective removal of transcription-blocking DNA damage from the transcribed strand of the mammalian DHFR gene. *Cell* **51,** 241–249 (1987).
3. Selby, C. P. & Sancar, A. Molecular mechanism of transcription-repair coupling. *Science* **260,** 53–58 (1993).
4. Savery, N. J. The molecular mechanism of transcription-coupled DNA repair. *Trends Microbiol.* **15,** 326–333 (2007).
5. Howan, K. *et al.* Initiation of transcription-coupled repair characterized at single-molecule resolution. *Nature* **490,** 431–434 (2012).
6. Graves, E. T. *et al.* A dynamic DNA-repair complex observed by correlative single-molecule nanomanipulation and fluorescence. *Nature Struct. Mol. Biol.* **22,** 452–457 (2015).
7. Deaconescu, A. M. *et al.* Structural basis for bacterial transcription-coupled DNA repair. *Cell* **124,** 507–520 (2006).
8. Westblade, L. F. *et al.* Structural basis for the bacterial transcription-repair coupling factor/RNA polymerase interaction. *Nucleic Acids Res.* **38,** 8357–8369 (2010).
9. Srivastava, D. B. & Darst, S. A. Derepression of bacterial transcription-repair coupling factor is associated with a profound conformational change. *J. Mol. Biol.* **406,** 275–284 (2011).
10. Park, J.-S., Marr, M. T. & Roberts, J. W. *E. coli* transcription repair coupling factor (Mfd protein) rescues arrested complexes by promoting forward translocation. *Cell* **109,** 757–767 (2002).
11. Smith, A. J., Szczelkun, M. D. & Savery, N. J. Controlling the motor activity of a transcription-repair coupling factor: autoinhibition and the role of RNA polymerase. *Nucleic Acids Res.* **35,** 1802–1811 (2007).
12. Haines, N. M., Kim, Y.-I. T., Smith, A. J. & Savery, N. J. Stalled transcription complexes promote DNA repair at a distance. *Proc. Natl Acad. Sci. USA* **111,** 4037–4042 (2014).
13. Strick, T. R., Allemand, J. F., Bensimon, D., Bensimon, A. & Croquette, V. The elasticity of a single supercoiled DNA molecule. *Science* **271,** 1835–1837 (1996).
14. Wang, M. D. *et al.* Force and velocity measured for single molecules of RNA polymerase. *Science* **282,** 902–907 (1998).
15. Revyakin, A., Liu, C., Ebright, R. H. & Strick, T. R. Abortive initiation and productive initiation by RNA polymerase involve DNA scrunching. *Science* **314,** 1139–1143 (2006).
16. Kou, S. C., Cherayil, B. J., Min, W., English, B. P. & Xie, X. S. Single-molecule Michaelis-Menten equations. *J. Phys. Chem. B* **109,** 19068–19081 (2005).
17. Manelyte, L., Kim, Y.-I. T., Smith, A. J., Smith, R. M. & Savery, N. J. Regulation and rate enhancement during transcription-coupled DNA repair. *Mol. Cell* **40,** 714–724 (2010).
18. Van Houten, B., Gamper, H., Sancar, A. & Hearst, J. E. DNase I footprint of ABC excinuclease. *J. Biol. Chem.* **262,** 13180–13187 (1987).
19. Selby, C. P. & Sancar, A. Structure and function of the (A)BC excinuclease of *Escherichia coli*. *Mutat. Res.* **236,** 203–211 (1990).
20. Taniguchi, Y. *et al.* Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329,** 533–538 (2010).
21. Revyakin, A., Ebright, R. H. & Strick, T. R. Promoter unwinding and promoter clearance by RNA polymerase: detection by single-molecule DNA nanomanipulation. *Proc. Natl Acad. Sci. USA* **101,** 4776–4780 (2004).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was made possible by a China Scholarship Council award to J.F., as well as grants from the French Agence Nationale pour la Recherche (RepOne) and the European Science Foundation (EURYI) to T.R.S., as well as core funding from the CNRS and the University of Paris Diderot. The Strick laboratory is also part of the Programme Equipe Labellisées of the Ligue Contre le Cancer. We thank N. Joly for assistance with protein purification and the Strick laboratory for feedback.

Author Contributions J.F., M.L.C., N.J.S. and T.R.S. planned experiments; J.F., M.L.C. and T.R.S. prepared reagents; J.F. performed tethered-RNAP, tethered-DNA and NanoCOSM assays; M.L.C. and T.R.S. performed tethered-RNAP assays. J.F., M.L.C. and T.R.S. conducted data analysis, and N.J.S. and T.R.S. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.R.S. (strick@biologie.ens.fr).

Reviewer Information *Nature* thanks S. Deindl, J. Elf and J. Roberts for their contribution to the peer review of this work.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

DNA constructs. Nanomanipulation constructs bearing the T5 N25 promoter[15] followed by either a C-less cassette (with the first C on the coding strand located at position +20 from the transcription start site, or TSS) or a CPD (at position +20 from the transcription start site, or TSS, and on the transcribed strand) were constructed as described previously[5,6], but with one modification. Specifically, enzyme reactions used to prepare the constructs (restriction and ligation) were not heat-inactivated, but rather were purified of protein by use of a spin column for DNA purification (Macherey-Nagel). The transcription unit is 5′-TTGCTTT CAGGAAAATTTTTCTGTATAATAAGCTTATAAATTTGAGAGAGGAGAC CAAATATGGCTGGTTCTCCACTAGTTCCGAATAG-3′, where the −35 and −10 promoter elements are underlined, the +1 TSS is in bold type, and the first C on which RNAP stalls at +20 is in bold type and underlined.

DNA nanomanipulation constructs used for tethered-RNAP translocation assays (as Fig. 1) were ~8 kbp long and contained the transcription unit at the DNA end distal to the surface and oriented in such a way as to direct initial transcription towards the surface[6].

DNA nanomanipulation constructs used for the tethered-DNA supercoiling assays (as Fig. 2) were all 2 kbp long and contained the transcription unit as a centrally located cassette, except for experiments in Extended Data Fig. 4 employing negatively supercoiled DNA bearing a C-less cassette, in which case a shortened, 1 kbp long DNA construct containing the transcription unit as a centrally located cassette was used for its enhanced spatial resolution[5,6,15,22].

DNA nanomanipulation constructs used for NanoCOSM assays (as Fig. 3) were all 3 kbp long and contained the transcription unit located ~400 bp from the coverslip surface and oriented so as to direct initial transcription towards the surface[6].

Proteins. *E. coli* RNAP, σ^{70}, GreB and Mfd, as well as SNAP-tagged RNAP (SNAP-RNAP) and SNAP-tagged Mfd (SNAP-Mfd) were purified as previously described[5,6]. Core RNAP was saturated with a threefold excess of σ^{70} to maintain polymerase in the holoenzyme form. SNAP-tagged proteins were labelled with BG-DY549 dye (New England Biolabs) as previously described[6]. UvrA, UvrB and UvrC proteins were purified via nickel-affinity chromatography as previously described[23], with the following modifications.

UvrA purified via nickel-affinity chromatography was then diluted sixfold with heparin buffer A (10 mM Tris·Cl pH 7.5, 50 mM KCl, 1 mM EDTA, 1 mM DTT, 5% glycerol) and loaded onto 10 ml of heparin resin (HiTrap Heparin, GE Healthcare) equilibrated in heparin buffer A. UvrA was eluted from the heparin resin by developing a gradient to 1 M KCl, concentrated as necessary to ~5 ml (10,000 MWCO Vivaspin 20, GE Healthcare) and then gel-filtrated (Superdex HiLoad 200 16/60, GE Healthcare) in GF buffer (10 mM TrisCl pH 8, 200 mM KCl, 1 mM EDTA, 2 mM DTT, 5% glycerol) before overnight dialysis into GF buffer containing 50% glycerol, aliquoting and snap-freezing in LN_2.

UvrB purified via nickel-affinity chromatography was similarly diluted sevenfold with heparin buffer A to a conductivity of ~6 mS cm^{-1} before being loaded onto 10 ml of heparin resin equilibrated in heparin buffer A and eluted by developing a gradient to 1 M KCl. Peak fractions were diluted to a conductivity of ~6 mS cm^{-1} using heparin buffer A, loaded onto 1 ml of anion exchange resin (MonoQ 5/50 GL, GE Healthcare) equilibrated in heparin buffer A, and eluted by developing a gradient to 1 M KCl. Peak fractions were pooled and concentrated as necessary to ~5 ml (10,000 MWCO Vivaspin 20, GE Healthcare) and then gel-filtrated (Superdex HiLoad 200 16/60, GE Healthcare) in GF buffer before overnight dialysis into GF buffer containing 50% glycerol, aliquoting and snap-freezing in LN_2.

UvrC purified via nickel-affinity chromatography was concentrated as necessary to ~5 ml (10,000 MWCO Vivaspin 20, GE Healthcare) and gel-filtrated (Superdex HiLoad 200 16/60, GE Healthcare) in GF buffer UvrC (10 mM TrisCl pH 7.5, 350 mM KCl, 1 mM EDTA, 5% glycerol, 2 mM DTT) before overnight dialysis into GF buffer UvrC containing 50% glycerol, aliquoting and snap-freezing in LN_2.

Protein concentrations were determined using the Folin–Lowry assay. Protein preparations were free of non-specific nuclease activity as determined by the absence of incision of supercoiled plasmid DNA in overnight reactions at 37 °C in standard repair buffer (RB; see below and refs 5, 6) and in large excess of protein. The fully reconstituted UvrABC system was observed to specifically nick ultraviolet-irradiated plasmid DNA as expected.

Glass surfaces. Surfaces used in single-molecule nanomanipulation assays were derivatized with anti-digoxigenin[22], and surfaces used in correlative NanoCOSM assays were derivatized with streptavidin[6].

Reaction conditions. Experiments were performed at 34 °C in repair buffer (RB) containing 40 mM K-HEPES pH 8.0, 100 mM KCl, 8 mM $MgCl_2$, 0.5 mg/ml BSA, 0.1% w/v Tween 20, and 10 mM β-mercaptoethanol (adapted from refs 15, 24). Unless specified otherwise, concentrations of components if present in reactions were as follows: 10–25 pM RNAP holoenzyme, 50 pM UvrA, 250 nM UvrB, 100 pM UvrC, 100 pM pUC18 competitor DNA, 500 nM GreB, 100 nM Mfd, 2 mM ATP and 200 μM each of UTP and GTP (for DNA bearing a C-less cassette) or 200 μM each of UTP, GTP and CTP (for DNA bearing a CPD). NanoCOSM assays were conducted at 27 °C. Collecting the specified number of events, *n*, typically requires more than three experimental runs involving usually 5–20 DNA molecules simultaneously (technical replicates).

Tethered-RNAP assays. Tethered-RNAP translocation assays (as Fig. 1) were performed as described previously[6]. In these assays a biotinylated RNAP is loaded at one end of an ~8 kbp DNA construct using the T5 N25 promoter, and stalled using the C-less cassette located at +20 from the TSS. The stalled RNAP–DNA construct is then bound to the standard streptavidin-coated magnetic beads used in all these assays (MyOne Streptavidin C1, Life Technologies). When deposited on an anti-digoxigenin-coated glass surface, the end of the ~8 kbp construct distal to the stalled RNAP, and which bears multiple digoxigenin groups, binds to the surface. The DNA is gently extended away from the surface using a low force ($F = 1$ pN). Displacement of stalled RNAP and formation of the translocating Mfd–RNAP complex is initiated by adding Mfd (100 nM) and ATP (2 mM).

Tethered-DNA assays. Tethered-DNA nanomanipulation experiments (as Fig. 2) were performed according to the procedures detailed in refs 15, 21, 22, 25. Standard reactions contained the following: extended, supercoiled DNA ($F = 0.3$ pN, superhelical density $\sigma = \pm 0.021$ for experiments using positive or negative supercoiled DNA, respectively), 10–25 pM RNAP holoenzyme, 500 nM GreB, 100 nM Mfd, and 2 mM ATP and 200 μM each of UTP and GTP (for DNA bearing a C-less cassette) or 2 mM ATP and 200 μM each of UTP, GTP and CTP (for DNA bearing a CPD).

Two methodologies—pulse-chase and continuous tracking—were used for these measurements. As previously shown[5,15] the methodologies are absolutely equivalent in terms of quantitative analysis and simply represent optimizations of experiments that can have characteristically long or short timescales, respectively, as detailed below.

Pulse-chase methodology. Single-round 'pulse-chase' assays[5], in which a single RNAP is stalled on DNA before free components are washed out and downstream components are flowed in, are optimal for the observation of long-lived repair intermediates, which must not be interrupted by reloading of a new RNAP molecule. Thus time-traces from pulse-chase experiments typically display a gap in the tracking data corresponding to the moment of component injection.

To stall RNAP on DNA we first injected 25 pM RNAP holoenzyme, 500 nM GreB and 200 μM nucleotides (ATP, UTP and GTP for experiments on DNA bearing a C-less cassette, and all four nucleotides for experiments on DNA bearing a CPD). Upon loading and stalling of RNAP, we wash out free RNAP holoenzyme while maintaining GreB and NTPs in solution. We then add 100 nM Mfd and 2 mM ATP to the reaction chamber to allow the reaction to begin with displacement of stalled RNAP.

This methodology was typically used to generate the quantitative kinetic data in which UvrAB are absent: Fig. 2b (time distribution) and Extended Data Figs 3 and 7c. This methodology was also used to preload RNAP for TCR incision rate measurements presented in Fig. 3c–e, in which case UvrA, UvrB and UvrC components were added as described below.

Continuous-tracking methodology. Continuous tracking assays[5], in which all components are simultaneously present in solution, are optimal for the statistical observation of short-lived repair intermediates which are only rarely interrupted by reloading of a new RNAP molecule. Here we injected 10–25 pM RNAP holoenzyme, 500 nM GreB, 2 mM ATP and 200 μM nucleotides (UTP and GTP for experiments performed using the C-less cassette, and UTP, GTP and CTP for experiments performed using a CPD), and UvrA and UvrB as specified.

This methodology was typically used to generate uninterrupted time-traces for presentation such as in Fig. 2b, c (time-traces) as well as the quantitative data shown in Figs 2c, d and 3a, b and Extended Data Figs 4–6 and 9.

NanoCOSM assays. NanoCOSM analysis is described in detail in ref. 6. Briefly, it is based on tethered-DNA experiments performed using a magnetic trap microscope into which a total internal reflection, or evanescent, field has been introduced[26]. For these assays combining topological measurement on torsionally constrained DNA and single-molecule fluorescence, a slightly longer DNA backbone (3 kbp) is employed to prevent the autofluorescent bead from entering too much into the evanescent field used to excite the fluorophore label. The transcription cassette is unchanged. DNA superhelical density is held constant at $|\sigma| = 0.021$, but for this longer DNA this corresponds to ±6 turns of the DNA in the magnetic trap rather than ±4 employed for 2 kbp DNA substrates. In these experiments, one fluorescent component was tested at a time. When used, SNAP-RNAP was at 50 pM and SNAP-Mfd was at 2.5 nM. All components other than the fluorescent one were present at the same concentrations as for standard tethered-DNA assays described

above, except that GreB was omitted and the bead and surface chemistries were inverted (that is, streptavidin-modified glass surface and anti-digoxigenin-coated magnetic bead[6]). NanoCOSM assays were performed using the continuous-tracking methodology.

UvrABC (GGR) incision assay. Incision by UvrABC of positively or negatively supercoiled DNA bearing either a C-less cassette or a CPD was performed in the presence of 1 nM UvrA, 250 nM UvrB, 100 pM UvrC, 2 mM ATP and 100 pM of competitor DNA (pUC18, used to reduce non-specific interactions between protein components and DNA).

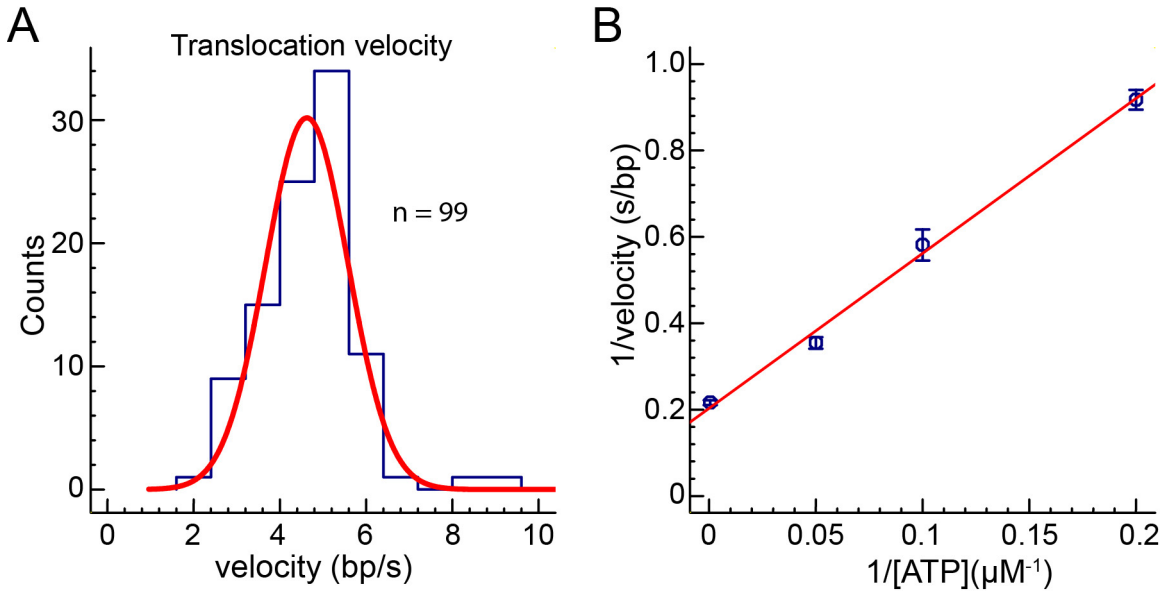
Incision by UvrABC of positively supercoiled DNA bearing a CPD protected by a stalled RNAP was performed by first stalling RNAP on nanomanipulated DNA (pulse-chase methodology as above) and then flowing in 1 nM UvrA, 250 nM UvrB, 100 pM UvrC, 2 mM ATP and 100 pM of competitor DNA.

Mfd–RNAP-UvrABC (TCR) incision assay. The TCR assay for positively or negatively supercoiled DNA bearing a CPD was performed using the single-round pulse-chase methodology described above. First, we stalled RNAP on the CPD by equilibrating the reaction cell with 20 pM RNAP, 200 μM of each of the four nucleotides, and 500 nM GreB. We then washed out free components while maintaining GreB and NTPs in solution. We then supplemented the reaction chamber with 1 nM UvrA, 250 nM UvrB, 100 pM UvrC, 2 mM ATP and 100 pM pUC18 competitor DNA, and then further supplemented the reaction chamber with 100 nM Mfd to displace stalled RNAP and initiate the TCR reaction.

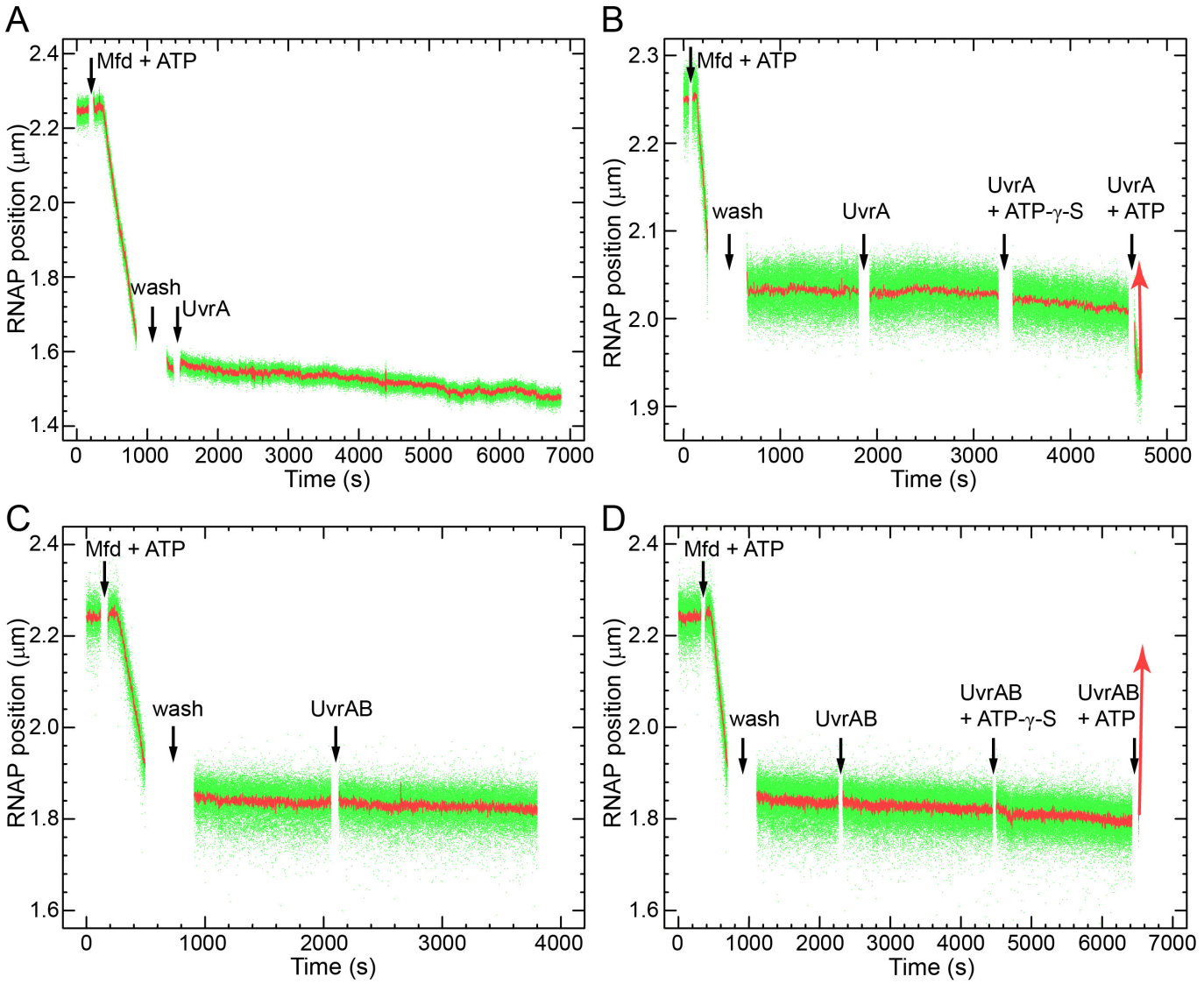
Data acquisition and analysis. Nanomanipulation data were collected on homebuilt magnetic traps running the Picotwist software suite for trap control and particle tracking and analysis (PicoTwist). Raw nanomanipulation data representing the magnetic bead position as observed under red illumination (650 nm) were collected at video rate (31 Hz, green points in time-traces) using a JAI CCD camera, and were filtered at ~1 s for analysis (red line in time-traces). Fluorescence data co-localized to the magnetic beads were collected using the Solis software suite provided by the EMCCD manufacturer (Andor) under 532 nm strobed illumination conditions (0.5 s illumination every 5 s), and were synchronized to the nanomanipulation data using dedicated timing trigger pulses generated by the programmable counters of the CCD camera.

Estimation of the fraction of Mfd–RNAP arrest events too short to be observed. To estimate the fraction of UvrA- or UvrAB-mediated Mfd–RNAP arrest events that are too short to observe, we first characterize the instrument response time. Thus the magnetic bead's vertical RMS fluctuations in the tethered-RNAP assay, where the extending force $F \approx 1$ pN, are ~15 nm. Given that the velocity of the Mfd–RNAP complex is only about 1.5 nm s^{-1}, we find an instrument response time of approximately 10 s. As events shorter than this may not be detectable, we estimate that the fraction of events that follow a single-exponential distribution with mean of 15 s, but have a duration shorter than 10 s, is of order 50%. Similarly, we estimate that the fraction of events that follow a single-exponential distribution with a mean of 6 s, but have a duration shorter than 10 s, is of order 80%. These values are in good agreement with observed fractions of missing arrest events, leading us to conclude that UvrA or UvrAB always arrest the translocating complex before dissociating it from the DNA.

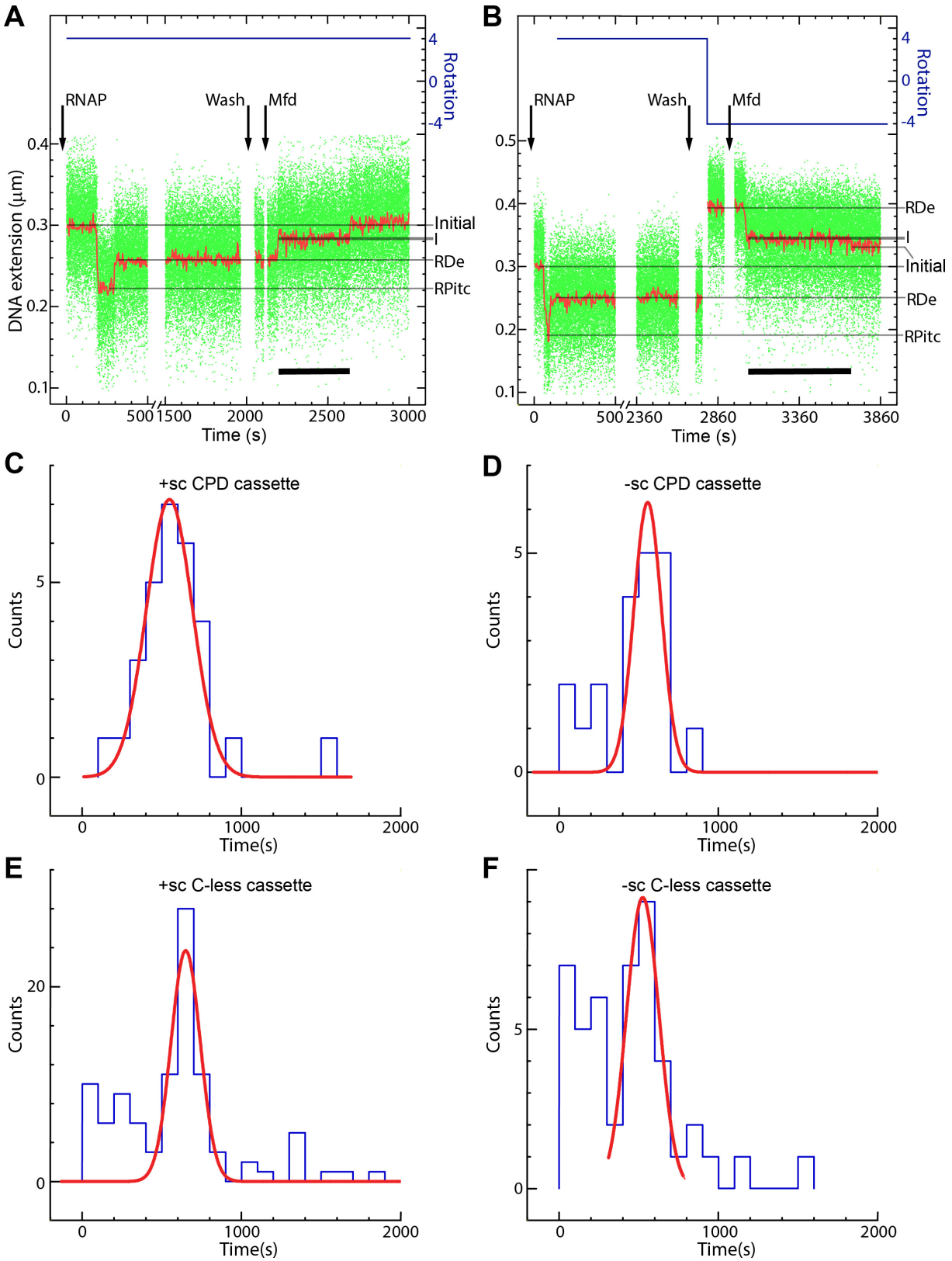
22. Revyakin, A., Ebright, R. H. & Strick, T. R. Single-molecule DNA nanomanipulation: improved resolution through use of shorter DNA fragments. *Nature Methods* **2,** 127–138 (2005).
23. Manelyte, L. *et al.* The unstructured C-terminal extension of UvrD interacts with UvrB, but is dispensable for nucleotide excision repair. *DNA Repair* **8,** 1300–1310 (2009).
24. Smith, A. J. & Savery, N. J. RNA polymerase mutants defective in the initiation of transcription-coupled DNA repair. *Nucleic Acids Res.* **33,** 755–764 (2005).
25. Revyakin, A., Allemand, J. F., Croquette, V., Ebright, R. H. & Strick, T. R. Single-molecule DNA nanomanipulation: detection of promoter-unwinding events by RNA polymerase. *Methods Enzymol.* **370,** 577–598 (2003).
26. Duboc, C., Graves, E. T. & Strick, T. R. Simple calibration of TIR field depth using the supercoiling response of DNA. *Methods* **105,** 56–61 (2016).
27. Verhoeven, E. E. A., Wyman, C., Moolenaar, G. F., Hoeijmakers, J. H. J. & Goosen, N. Architecture of nucleotide excision repair complexes: DNA is wrapped by UvrB before and after damage recognition. *EMBO J.* **20,** 601–611 (2001).
28. van den Broek, B., Noom, M. C. & Wuite, G. J. L. DNA-tension dependence of restriction enzyme activity reveals mechanochemical properties of the reaction pathway. *Nucleic Acids Res.* **33,** 2676–2684 (2005).



Extended Data Figure 1 | Motor properties of the translocating Mfd–RNAP complex as seen in the tethered-RNAP assay. a, Velocity distribution of translocating Mfd–RNAP in the presence of 2 mM ATP and under a weak opposing load ($F = 1$ pN). Velocity is measured by fitting a single line segment to an entire translocation time-trace; this is made possible by the fact that velocity is essentially constant over the ~7,000 bp of displacement which constitutes an entire trajectory. The velocity distribution is fitted to a Gaussian, giving a mean velocity of 4.7 ± 1 bp s^{-1} (SD; $n = 99$ trajectories). **b**, Tau plot of inverse velocity of translocating Mfd–RNAP as a function of inverse ATP concentration is well-fitted to a line, indicating Michaelian behaviour with $K_m^{ATP} = 16 \pm 0.4\,\mu M$ (s.e.m.) and $V_{max}^{ATP} = 4.7 \pm 0.1$ bp s^{-1} (s.e.m.). Error bars, s.e.m. as determined from at least ten trajectories for each ATP concentration.

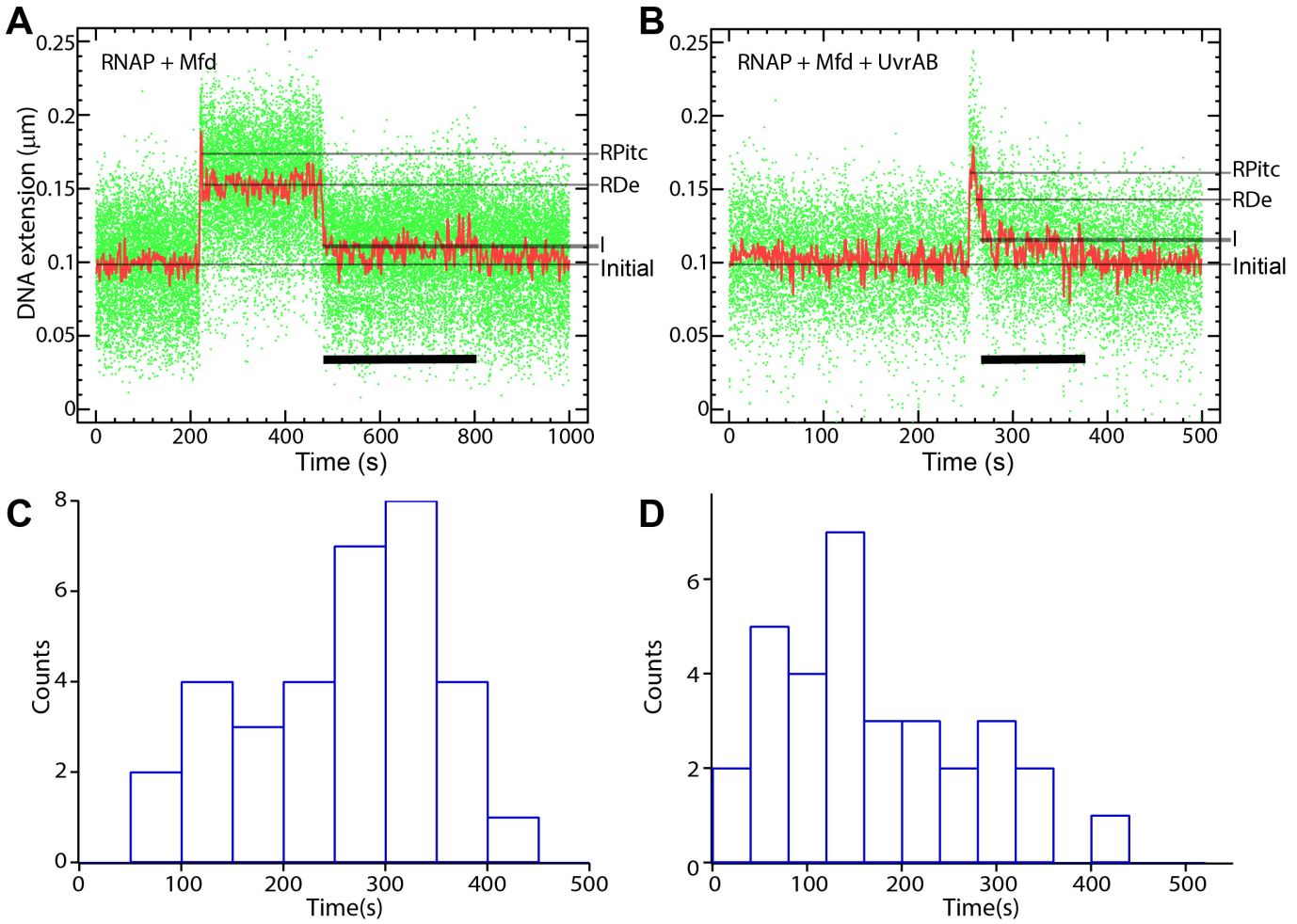


Extended Data Figure 2 | Resolution of the translocating Mfd–RNAP complex by UvrA or by UvrAB is ATP-dependent as shown by the tethered-RNAP translocation assay. Down-arrows indicate addition of components as noted and as follows. Beginning with stalled RNAP, we add 100 nM Mfd and 2 mM ATP to form the translocating Mfd–RNAP complex. A wash step using 5 ml of reaction buffer lacking ATP is applied to remove (nearly) all the free ATP in solution, causing the translocating complex to come to a nearly complete halt. Then, (**a**, **b**) 50 pM UvrA or (**c**, **d**) 50 pM UvrA and 250 nM UvrB is added to the experiment. The complex is stable and release of the magnetic bead is not observed. Further addition of ATP-γ-S (2 mM; see **b**, **d**) does not permit bead release. However, final addition of ATP (2 mM) leads to rapid release. Red up-arrows indicate bead release in **b**, **d**.

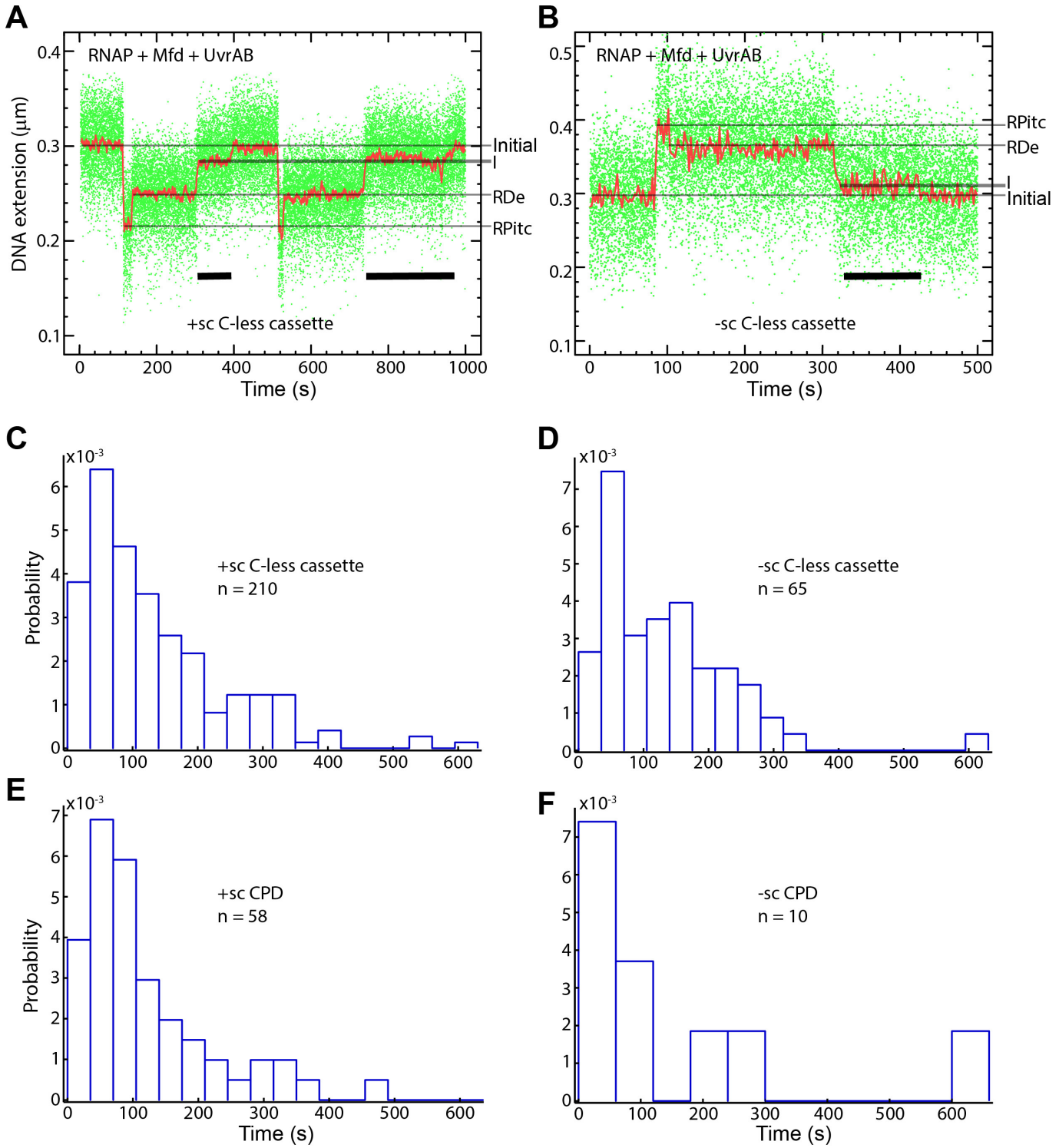


Extended Data Figure 3 | See next page for caption.

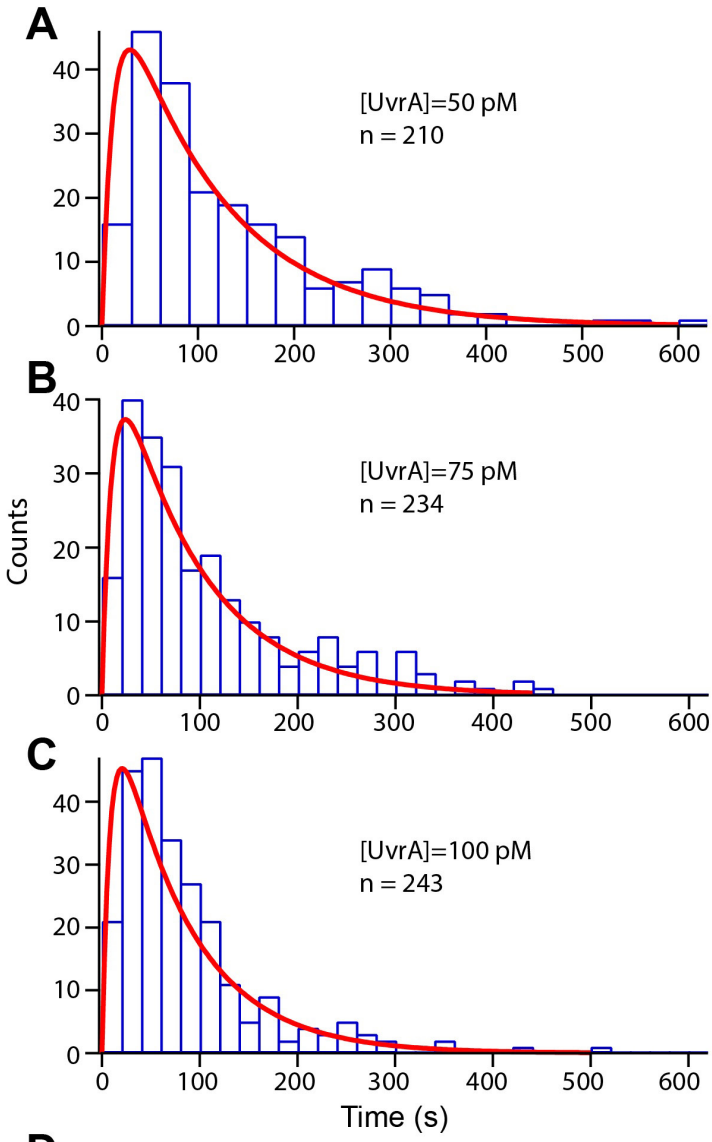
Extended Data Figure 3 | Characterization of the long-lived Mfd–RNAP intermediate on 2 kb DNA using the tethered-DNA assay. **a**, **b**, Nanomanipulation time-traces showing pulse-chase measurement of the lifetime of the Mfd–RNAP intermediate for CPD-bearing DNA under conditions of positive (+sc) and negative (−sc) supercoiling, respectively. Down-arrows indicate moments of component addition as noted and as follows. First, we load RNAP onto DNA in standard conditions (25 pM RNAP holoenzyme, 500 nM GreB, and the appropriate nucleotide complement, each present at 200 μM). We then wash out free RNAP with reaction buffer supplemented with 500 nM GreB and the nucleotide complement. We then initiate formation of the intermediate by infusion of the above wash solution supplemented with 100 nM Mfd and 2 mM ATP. For negatively supercoiled DNA, RNAP was loaded under conditions of positive supercoiling before the DNA was returned to negative supercoiling; blue line indicates when DNA supercoiling is changed. Black bar highlights the intermediate state. **c**, **d**, Lifetime distributions for the Mfd–RNAP intermediate formed on CPD-bearing DNA under conditions of positive or negative supercoiling, respectively, are well fitted to Gaussian distributions (red lines). For positive supercoiling the mean lifetime of the repair intermediate is 548 ± 37 s (s.e.m., $n = 29$ events: this distribution is also presented in Fig. 2b), and for negative supercoiling the mean lifetime of the repair intermediate is 556 ± 33 s (s.e.m., $n = 21$ events). **e**, **f**, Lifetime distributions for the Mfd–RNAP intermediate formed on C-less cassette DNA under conditions of positive or negative supercoiling, respectively, are also well fitted to Gaussian distributions. For positive supercoiling the mean lifetime of the repair intermediate is 649 ± 13 s ($n = 98$ events) and for negative supercoiling the mean lifetime of the repair intermediate is 524 ± 26 s ($n = 46$ events). Data were fitted over the range delimited by the red lines.



Extended Data Figure 4 | Characterization of the Mfd–RNAP intermediate on negatively supercoiled 1 kb DNA in the absence or presence of UvrAB using the tethered-DNA assay. DNA bears a C-less cassette. **a**, Nanomanipulation time-trace obtained in continuous-tracking mode in the presence of 25 pM RNAP holoenzyme, 100 nM Mfd, 500 nM GreB, 2 mM ATP, 200 μM UTP and 200 μM GTP. **b**, As in **a** but in the added presence of 50 pM UvrA and 250 nM UvrB. **c**, Lifetime distribution of the Mfd–RNAP intermediate in the absence of UvrA and UvrB has a mean lifetime of 258 ± 17 s (s.e.m., $n = 33$ events). **d**, Lifetime distribution of the Mfd–RNAP intermediate in the added presence of 50 pM UvrA and 250 nM UvrB has a mean lifetime of 167 ± 17 s (s.e.m., $n = 32$ events). Data sets for kinetics were obtained using the pulse-chase methodology.



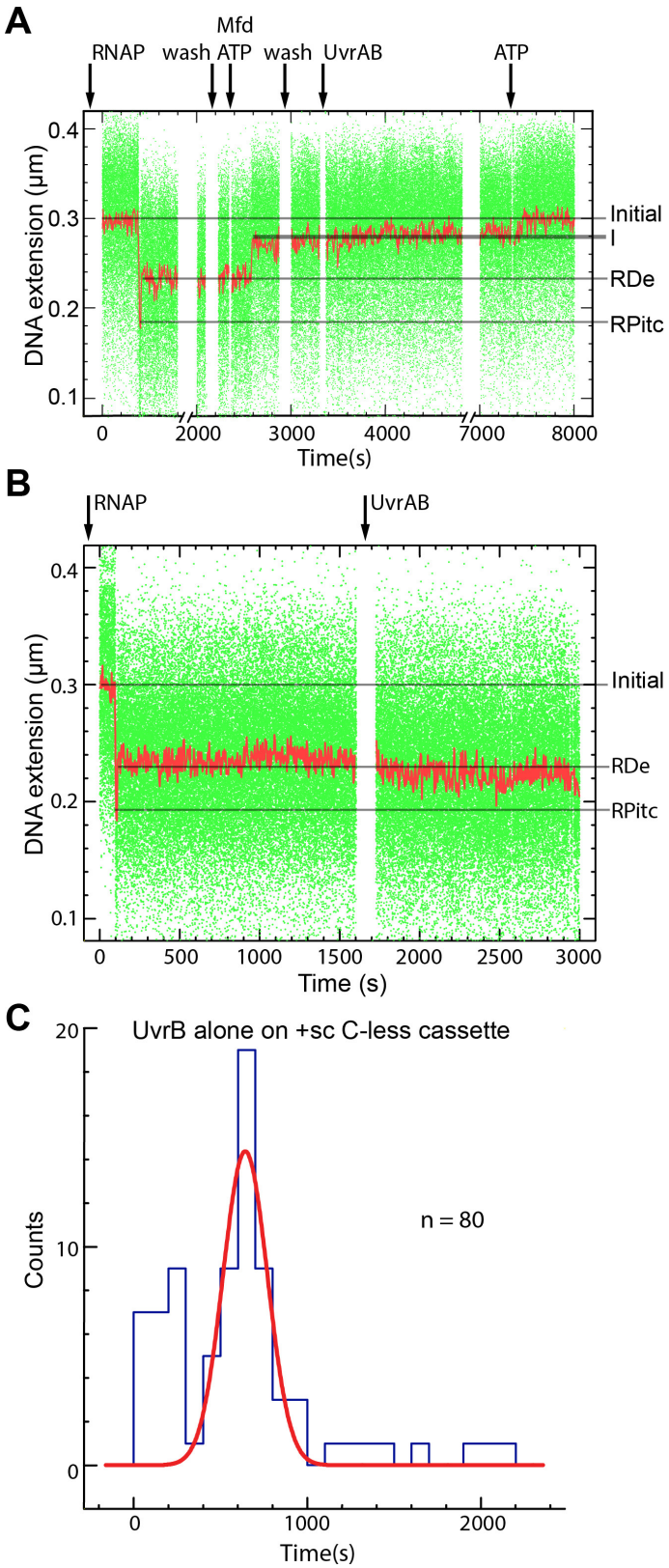
Extended Data Figure 5 | Reduced lifetime of the Mfd–RNAP intermediate in the presence of UvrAB, monitored on 2 kb DNA using the tethered-DNA assay. a, Nanomanipulation time-trace for positively supercoiled DNA (+sc) bearing a C-less cassette in the presence of 25 pM RNAP holoenzyme, 100 nM Mfd, 50 pM UvrA, 250 nM UvrB, 500 nM GreB, 2 mM ATP, 200 μM UTP and 200 μM GTP (continuous-tracking methodology). **b**, Nanomanipulation time-trace obtained as in **a** but for negatively supercoiled DNA (−sc). **c**–**f**, Lifetime distributions for the Mfd–RNAP intermediate in the presence of UvrA and UvrB as above are essentially independent of both the cause of RNAP stalling (either a C-less cassette or a CPD) and supercoiling of the DNA (positive or negative). For positively supercoiled template the mean lifetime observed using DNA bearing a C-less cassette is 132 ± 7 s (s.e.m., $n = 210$ (**c**), see overview in Extended Data Fig. 6d) and for DNA bearing a CPD it is 141 ± 20 s (s.e.m., $n = 58$ (**e**)). For negatively supercoiled template the mean lifetime observed using DNA bearing a C-less cassette is 132 ± 13 s (s.e.m., $n = 65$ (**d**)) and for DNA bearing a CPD it is 157 ± 65 s (s.e.m., $n = 10$ (**f**)).



D

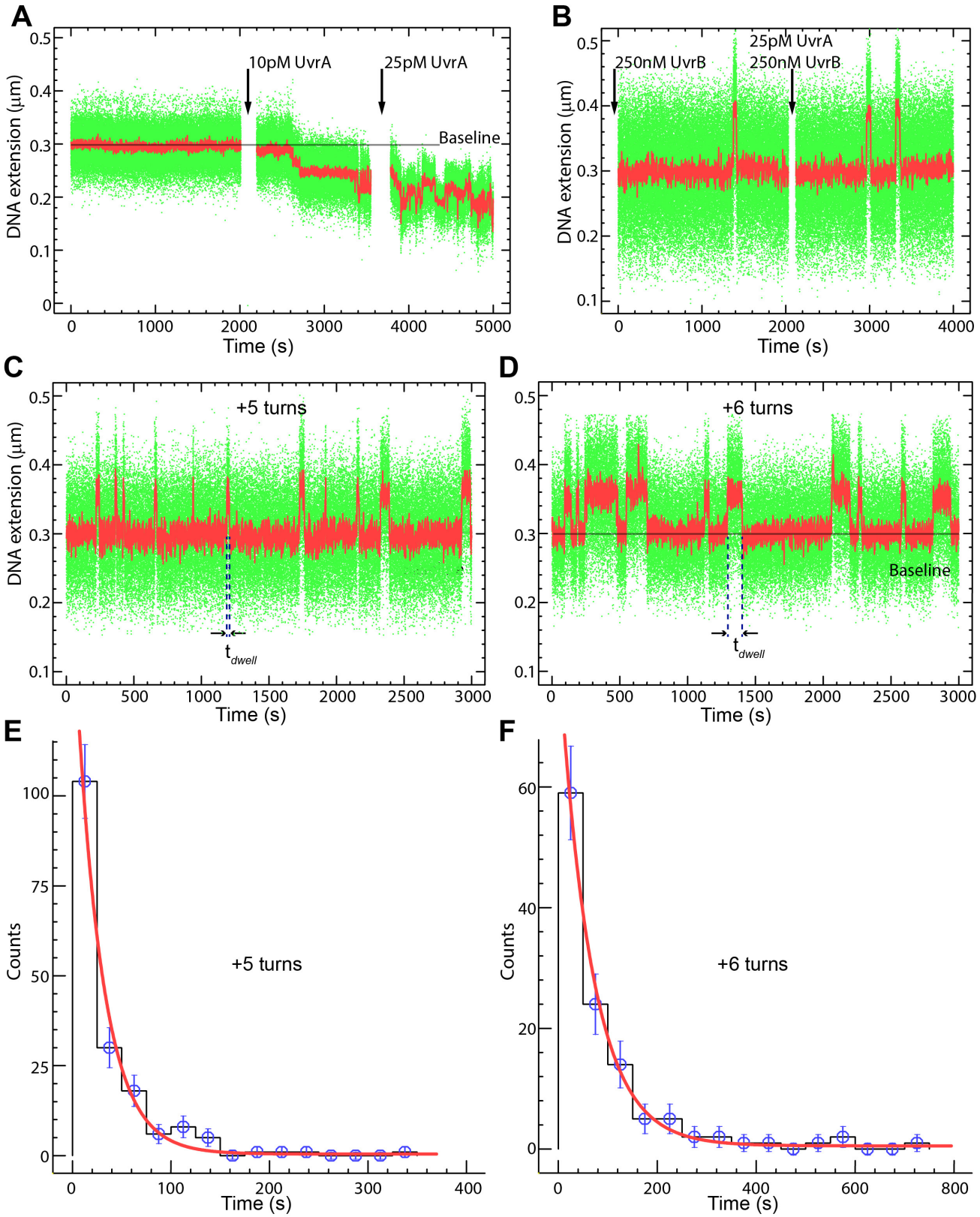
Supercoiling	DNA	[UvrA] (pM)	Mean lifetime of inter-mediate(s)	Number of events	σ	S.E.M.(s)
Positive supercoiling	C-less cassette	50	132	210	108	7
		75	110	234	94	6
		100	86	243	75	5
	CPD cassette	50	141	58	153	20
		100	85.7	52	70.7	9.8
Negative supercoiling	C-less cassette	50	132	65	103	12.7
	CPD cassette	50	157	10	205	64.8

Extended Data Figure 6 | Lifetime distributions of the Mfd–RNAP intermediate as a function of UvrA concentration, using the tethered-DNA assay. The DNA substrate used in these experiments was positively supercoiled and contained a C-less cassette, and data were collected using the continuous-tracking methodology in the presence of 10–20 pM RNAP holoenzyme, 500 nM GreB, 100 nM Mfd, 2 mM ATP, 200 μM UTP, 200 μM GTP and 250 nM UvrB. The UvrA concentration was (**a**) 50 pM, (**b**) 75 pM and (**c**) 100 pM. Red lines show the result of global fitting to a difference-of-two exponentials characteristic of a Michaelis–Menten process, using the rate-limiting forward catalytic step extracted from classical Michaelian analysis of the mean times (Fig. 2d) as an additional constraint. **d**, Overview of lifetimes of the Mfd–RNAP complex measured with the tethered-DNA assay and as a function of template supercoiling, cause of RNAP stalling and UvrA concentration, as presented in this paper. UvrB was fixed at 250 nM throughout.



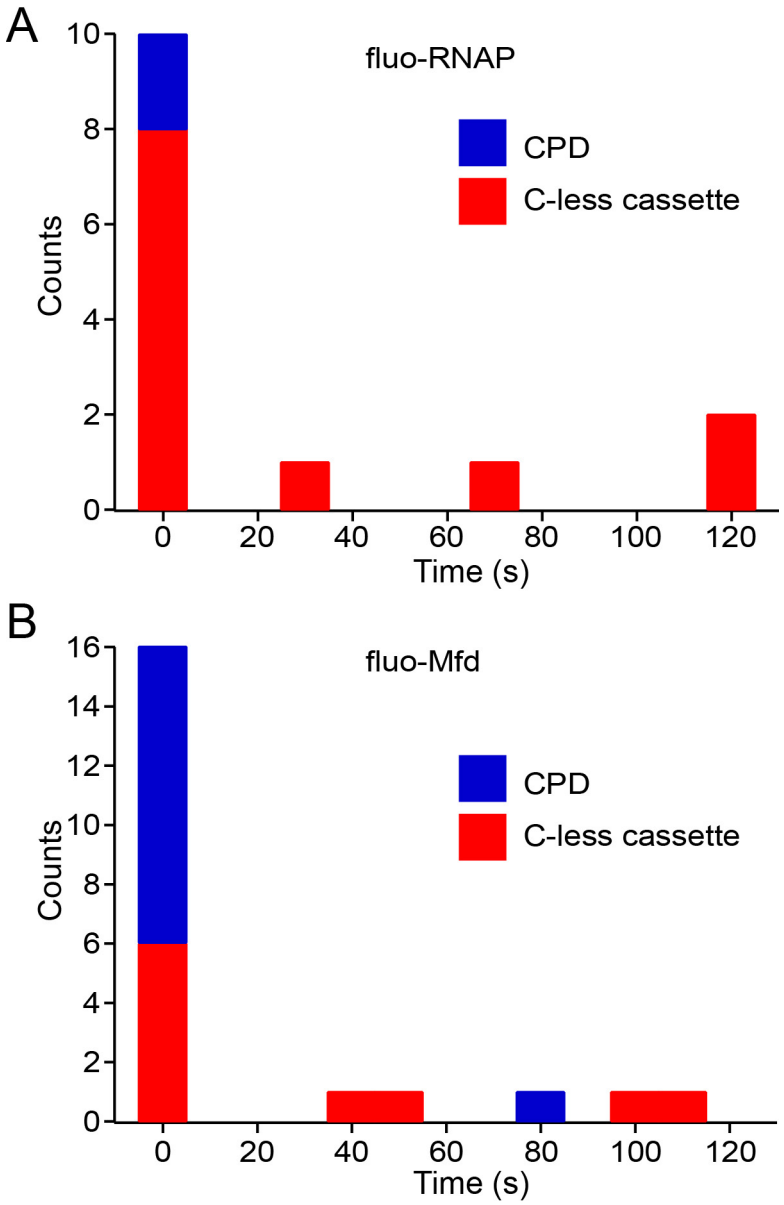
Extended Data Figure 7 | See next page for caption.

Extended Data Figure 7 | Mfd–RNAP-UvrAB control experiments, using the tethered-DNA assay. These experiments were performed on positively supercoiled DNA substrate (+sc) bearing a C-less cassette and using the standard pulse-chase methodology. **a**, No ATP control: ATP dependence of UvrAB remodelling of Mfd–RNAP intermediate. Black down-arrows indicate component infusion as follows. RNAP: we first introduce 25 pM RNAP holoenzyme, 500 nM GreB, 200 μM ATP, 200 μM UTP and 200 μM GTP, and wait for RNAP to stall on DNA. Wash: we next wash out all free components except for GreB. Mfd ATP: we next infuse 100 nM Mfd, 500 nM GreB and 50 μM ATP and wait for Mfd to remodel RNAP and form the Mfd–RNAP intermediate. Wash: we next wash out all free components except GreB. UvrAB: we next infuse 50 pM UvrA, 250 nM UvrB and 500 nM GreB. We wait several thousand seconds, without any observed change in the intermediate state. ATP: finally, we infuse 2 mM ATP into the reaction and rapidly observe resolution of the intermediate species. **b**, No Mfd control: UvrAB does not functionally interact with RNAP in the absence of Mfd. Stalled RNAP formed as in **a** is not displaced in the presence of (down-arrow) 50 pM UvrA, 250 nM UvrB, 500 nM GreB and 2 mM ATP. **c**, No UvrA control: lifetime distribution for the Mfd–RNAP intermediate in the presence of UvrB alone. Stalled RNAP is formed as in **a**. We then wash out free RNAP while maintaining GreB and NTPs in solution, and add 100 nM Mfd, 250 nM UvrB and 2 mM ATP while maintaining GreB and NTPs in solution. The lifetime of the Mfd–RNAP intermediate thus formed remains long-lived (642 ± 22 s (s.e.m.), $n = 80$) with Gaussian statistics.

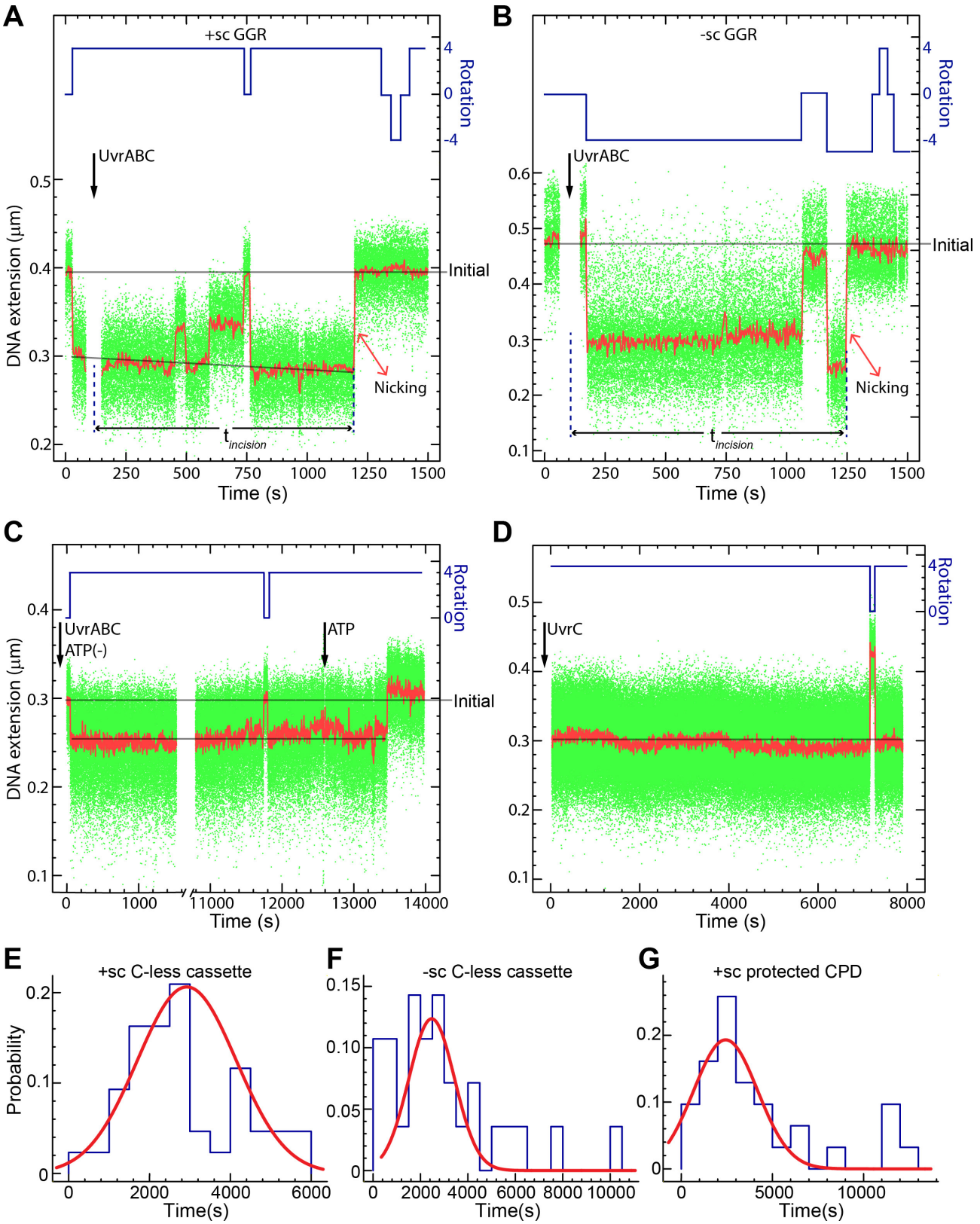


Extended Data Figure 8 | See next page for caption.

Extended Data Figure 8 | Control experiments for UvrA and UvrB interactions with DNA in the absence of damage as seen in the tethered-DNA supercoiling assay. Experiments were conducted on positively supercoiled DNA bearing a CTP-less cassette. **a**, UvrA alone compacts undamaged, supercoiled DNA in a non-specific manner even at concentrations as low as 10 pM. Trace shown obtained with 1 mM ATP; the same phenomenon is observed in the absence of ATP (data not shown). **b**, UvrB prevents non-specific interaction of UvrA with DNA; ($t = 0$ s) 250 nM UvrB alone does not compact DNA, although it transiently interacts non-specifically and briefly with DNA in the presence of 1 mM ATP (see **c**–**f**). The same phenomenon is observed in the absence of ATP (data not shown); ($t = 2{,}000$ s) addition of UvrB also prevents UvrA from compacting DNA non-specifically. On the basis of these data we set the working UvrB concentration to 250 nM: our measurements with UvrA can thus go up to 100 pM, which remains more than 90% saturated by this concentration of UvrB as shown by the fact that we can perform measurements without DNA compaction. **c**, **d**, Time-traces obtained on positively supercoiled DNA in the presence of 250 nM UvrB and 1 mM ATP show supercoiling-dependence of the dwell time (t_{dwell}) of UvrB-DNA 'wrapping' events. Indeed the amplitude of these events (~50–100 nm) is consistent with titration of a large positive supercoil by formation of a tight/compact, positive wrap of DNA around UvrB as observed in AFM imaging[27]. **e**, **f**, Histograms of the dwell time of the wrap state obtained above are fitted to single-exponential distributions, with a mean dwell time of (**e**) 28 ± 2 s (s.e.m., $n = 175$, +5 turns), and (**f**) 66 ± 5 s (s.e.m., $n = 117$, +6 turns). By performing experiments with no more than 250 nM UvrB and with only +4 turns of positive supercoiling, this wrap state is of order 10 s and does not significantly interfere with detection of Mfd–RNAP intermediates or their resolution, and UvrB safely inhibits DNA compaction activity by UvrA.



Extended Data Figure 9 | Correlation between resolution of the Mfd–RNAP intermediate and loss of fluorescent signal from labelled RNAP or Mfd in the NanoCOSM assay. We plot the time elapsed between loss of fluorescence signal from (**a**) fluorescent RNAP or (**b**) fluorescent Mfd and nanomechanical resolution of the Mfd–RNAP intermediate as observed in the magnetic trap, as shown in Fig. 3a, b. In both cases the vast majority of events are correlated as shown by the fact that loss of fluorescence and nanomechanical resolution of the intermediate temporally coincide (that is, the time between the two events is nil). Loss of fluorescence before nanomechanical resolution (that is, indicated as positive times) is most probably due to spontaneous photobleaching of the DY-549 fluorophore used to label proteins. No significant difference is observed between DNA substrates bearing a CPD or bearing a C-less cassette.



Extended Data Figure 10 | See next page for caption.

Extended Data Figure 10 | Characterization of specific, control and non-specific GGR incision using the tethered-DNA assay. a, **b**, Time-traces showing GGR incision on positively and negatively supercoiled CPD-bearing DNA. For positively supercoiled DNA (+sc), addition of UvrABC proteins (1 nM UvrA, 250 nM UvrB, 100 pM UvrC and 100 pM pUC18 competitor DNA) and 1 mM ATP led to DNA incision and an abrupt loss of supercoiling. The average GGR incision times are 1,230 ± 195 s (s.e.m., $n = 72$ events) and 1,156 ± 256 s (s.e.m., $n = 40$ events) for +sc and −sc, respectively; see Fig. 3f, g for distributions and fits. **c**, As in **a**, but in the absence of ATP. The absence of incision was confirmed on 22 molecules over a ~4 h window. Upon supplementing the reactions with 1 mM ATP (red down-arrow) incision rapidly takes place. **d**, UvrC (100 pM) and ATP (1 mM) are unable to incise positively supercoiled, CPD-bearing DNA. The absence of incision was confirmed on 31 molecules over a ~2 h window. **e**, Incision times for UvrABC (as above) acting on positively supercoiled DNA bearing a C-less cassette (that is, undamaged) are essentially normally distributed (red line) with a mean of 2,922 ± 222 s ($n = 44$ events; the fit was obtained by excluding points between 3,000 and 4,000 s). **f**, Incision times for UvrABC (as above) acting on negatively supercoiled DNA bearing a C-less cassette are essentially normally distributed with a mean of 2,471 ± 377 s (s.e.m., $n = 28$ events; the fit was obtained by excluding points below 1,000 s). **g**, Incision times for UvrABC acting on positively supercoiled DNA bearing a CPD protected by stalled RNAP in the absence of Mfd are essentially normally distributed with a mean of 2,348 ± 672 s (s.e.m., $n = 31$ events). Red lines are guides to the eye (**c**, **d**) and overall results confirm all of UvrAB, UvrC and ATP are required for GGR incision. Results from **e**–**g** further indicate that non-specific DNA incision by the complete GGR system can take place in this assay; however, it is slow enough to permit measurement of faster specific incision rates discussed in Fig. 3. We propose these incision events are in fact specific to the multiple biotin- and digoxigenin-based tethers at the ends of the DNA construct, which can ultimately be recognized as DNA damage by the GGR machinery[9]. Because only nicking at the first tethering biotin or dig, and within the 2 kbp fragment, will result in loss of supercoiling, then, statistically, multiple incisions at multiple tethers must be realized before loss of supercoiling, resulting in a normal distribution. This can be compared to DNA incision by endonuclease, the time distribution of which is single-exponential[28]. As incision on these constructs is significantly slower and obeys different statistics than that on CPD-containing DNA, we conclude that our single-molecule measurements can indeed isolate CPD-specific from non-specific incision by the GGR machinery.

CORRIGENDUM

doi:10.1038/nature18000

Corrigendum: Dietary emulsifiers impact the mouse gut microbiota promoting colitis and metabolic syndrome

Benoit Chassaing, Omry Koren, Julia K. Goodrich, Angela C. Poole, Shanthi Srinivasan, Ruth E. Ley & Andrew T. Gewirtz

Nature **519,** 92–96 (2015); doi:10.1038/nature14232

Some clarifications are provided to this Letter; these do not alter any of the central conclusions but, rather, are provided in the interests of transparency and reproducibility. Our Letter indicated that experiments were performed on 4-week-old mice (unless stated otherwise). In fact, for several experiments, mice ranged from 5 to 7 weeks as follows: Fig. 4a–h, Extended Data Fig. 9g–w, b′–t′, z: 5 weeks old; Figs 1, 2, 3a–d, 4i–o, Extended Data Figs 1a–d, 2, 4, 5s–v, 6, 7h–k, 8l–s, 9a–f, x, y, 10: 6 weeks old; Fig. 3e–l, Extended Data Fig. 1e–l, 5q, r, 7a–g, l–h′, 8a–k, t–o′, 9a′: 7 weeks old. The weight gain versus time curves are affected by mouse age and hence explain why the kinetics of weight gain differ among control mice when comparing between different experiments.

For each experiment, we listed the average *n* value for all the conditions within each panel, which differed from the exact *n* for each experimental condition within each figure, as shown in Supplementary Table 1 of this Corrigendum.

Furthermore, our Letter reported relative changes in body mass and absolute mass of fat pads, thus not permitting assessment of absolute weight changes nor fat pad mass relative to total body mass. Hence, we provide measures of absolute and relative body and fat pad mass in a side-by-side manner in the Supplementary Data to this Corrigendum.

Supplementary Information is available in the online version of this Corrigendum.

CORRECTIONS & AMENDMENTS

CORRIGENDUM

doi:10.1038/nature17982

Corrigendum: Kidney organoids from human iPS cells contain multiple lineages and model human nephrogenesis

Minoru Takasato, Pei X. Er, Han S. Chiu, Barbara Maier, Gregory J. Baillie, Charles Ferguson, Robert G. Parton, Ernst J. Wolvetang, Matthias S. Roost, Susana M. Chuva de Sousa Lopes & Melissa H. Little

Nature **526,** 564–568 (2015); doi:10.1038/nature15695

In the Methods of this Letter, '5,000' should have read '15,000' in the sentence: "Then, cells were again plated on a Matrigel-coated at 15,000 cells per cm^2 in MEF-CM." This error has been corrected online.

CORRECTIONS & AMENDMENTS

ERRATUM

doi:10.1038/nature18019

Universal resilience patterns in complex networks

Jianxi Gao, Baruch Barzel & Albert-László Barabási

Nature **530,** 307–312 (2016); doi:10.1038/nature16948

In the last sentence of page 310 of this Letter, the parameter h should equal 2, rather than 1. In addition, after equation (4), the text should have stated '$A_{ij} > 0$' and 'positive interactions', to read "...the weighted connectivity matrix $A_{ij} > 0$ captures the positive interactions between the nodes.". These errors have been corrected online.

AERONAUTICS A schoolgirl space enthusiast grows up to rescue a spacecraft **p.241**

ENTREPRENEURSHIP How to get a start-up company started **go.nature.com/2actuzb**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

PHILIP LEE HARVEY/CULTURA RM EXCLUSIVE/GETTY

Ecology is one of a few fields moving towards the multiple-working-hypotheses method of investigation.

RESEARCH PROTOCOLS

A forest of hypotheses

Falling in love with a single theory can cut off fruitful avenues of enquiry. Here's how to keep your mind open.

BY JULIA ROSEN

The clamour in a Panamanian rainforest is deafening to human ears: bugs shriek, birds sing and bats screech throughout the humid night. To avoid attracting predators, male katydids (*Tettigoniidae*) trill out short, infrequent mating calls less than a second long.

Postdoc Laurel Symes, who studies sensory perception and decision-making at Dartmouth College in Hanover, New Hampshire, wants to understand how female katydids find their mates. She first thought they must have highly sensitive hearing. But she juggles other ideas at the same time: maybe katydids always meet up on a certain type of host plant, have neural mechanisms that filter out background noise or use another trick entirely.

These aren't just idle musings: Symes's collection of hypotheses is an integral part of her research. The approach helps her to home in on answers and avoid investment in a sole idea — a common tendency in science that can lead to trouble. History contains numerous examples of scientists who missed important clues because they clung too tightly to a favourite hypothesis. One way to avoid this fate is to consider many potential hypotheses.

Proponents of the multiple-working-hypotheses method say that it prevents scientists from developing 'tunnel vision', and enables them to embrace the possibility that several hypotheses might be true at once. Practising the approach takes discipline: researchers must brainstorm possible explanations for a scientific phenomenon before collecting or analysing data, and use techniques such as scrambling the order of samples and blinding data to help to counteract favouritism. It also demands that scientists remain open-minded during the entire research process, and continually refine their hypotheses.

A LONG HISTORY

The method of multiple working hypotheses was formally articulated[1] in 1890 by geologist Thomas Chrowder Chamberlin, then president of the University of Wisconsin–Madison. Building on the ideas of fellow geologist Grove Karl Gilbert, Chamberlin warned that when scientists come up with an original idea, they tend to develop affection for it, which can cloud their ability to do objective work. He argued that the solution was to generate and explore a family of hypotheses. By coming up with alternatives, he suggested, scientists would not be inclined to favour one idea.

Although the concept has faced criticism, aimed mainly at the impossibility of conceiving — let alone testing — all possibilities, many scientists say that it is as relevant today as ever. The pressure to publish in high-profile journals, win grants and build a reputation can prompt researchers — consciously or not — to seek support for pet ideas. One study posted to the preprint server arXiv[2] in June found that when programmers introduced these kinds of incentives into a model, simulated research groups succumbed to pressures to show support for original ideas, often erroneously.

Ecologist Barry Brook of the University of Tasmania in Australia thinks that resurrecting Chamberlin's ideas could help. In 2007, he co-authored a paper on the merits of using multiple working hypotheses for twenty-first-century science[3]. In many cases, he argues, the method produces more insightful results than testing null hypotheses, which reveals only whether a specific factor has a discernible effect. Multiple hypotheses, by contrast, can help scientists ▶

to work out whether that effect is important, and whether several factors might be at play.

Brook, for example, wanted to know why small mammals such as brown bandicoots (*Isoodon macrourus*) were disappearing from Northern Australia's Kakadu National Park. Many scientists had pointed in the past to introduced predators, such as cats, which seemed plausible. But when he considered other hypotheses and looked at historical population data, he found that cats had a negligible role, and that intense wildfires bore most blame[4]. "You can be surprised at how little support most of your well-crafted hypotheses can have," he says.

It might seem simpler to consider just one possible explanation, but ignoring other models can be dangerous. "That's not only dishonest, but it will also lead you down bad inferential pathways," Brook says.

RESIST TEMPTATION

It can be challenging to put the method into practice because researchers must battle their own natural enthusiasm for an alluring idea. The first step is to set aside time to articulate other hypotheses before one starts to gain traction. If not, a favoured hypothesis might skew the process of data collection or analysis when one heads out into the field, starts an experiment or dives into a data set. "If you have a hypothesis or you're looking for a pattern, sometimes you won't actually honour what pattern is there," says Kathleen Nicoll, a geographer at the University of Utah in Salt Lake City.

When coming up with a collection of hypotheses, it can be helpful to have patience and consult labmates — and to include a seemingly outrageous hypothesis. This idea was first advocated in 1926 by William Morris Davis, a retired geologist from Harvard University in Cambridge, Massachusetts, as a way to break out of conventional thinking. Many notable scientific advances fall into this category, including Alfred Wegener's then-scandalous claim in 1912 that continents migrate across Earth's surface[5] (they do), and the heretical proposal, developed in the 1920s by geologist J Harlen Bretz, that a catastrophic flood scoured out the heavily channeled landscapes of Washington (in fact, many violent floods swept through the region).

Symes finds that using multiple hypotheses yields the best results if researchers generate ideas that rely on different processes and make distinct predictions. In her research, a host-plant preference might lead to katydids having the same food in their guts, whereas using sound might imply that female katydids in Panama have more sensitive ears than species in forests without predatory bats. By identifying possible outcomes, she can design her experiments in ways that help to distinguish these ideas. "If the hypotheses are mutually exclusive or different in their mechanism, then you are going to learn something," she says.

Consideration of multiple working hypotheses continues during data processing and analysis, when scientists must take other steps to protect their objectivity (see 'Don't play favourites').

For Lydia Tackett, who studies marine fossils at North Dakota State University in Fargo, the solution is as simple as analysing samples out of order. Working chronologically through a geological sequence led her to identify trends prematurely and anticipate what she would find in subsequent layers. "Now, I collect the bulk samples I need and randomize the order," she says. She codes them so that she doesn't know exactly which layer each sample came from.

Others rely on statistical tools. Instead of using *P* values to reject individual models one at a time, Trevor Branch, a fisheries scientist at the University of Washington in Seattle, embraces a model-selection technique called Akaike's information criterion (AIC). This statistical method determines which of a set of models best explains data collected about an often-complex system. Branch says that it's a mathematical way of implementing Chamberlin's method of multiple working hypotheses.

Brook uses the AIC as well as the similar Bayesian information criterion, which is useful for distinguishing between a few simple models. When several models seem to be true, these methods help to weight their relative importance, so that their combined effects can be explored through something called multimodel inference. That involves merging several different models and considering them simultaneously to explain as much as possible.

Physicists and astronomers often take extreme measures to prevent researcher bias from creeping into their analyses. Saul Perlmutter, an astrophysicist at the University of California, Berkeley, relies on software or colleagues to hide potentially telling clues in the data before he sees them, a technique called blind analysis. This might include adding randomly generated numbers to data values, shifting them by random amounts or hiding the axes on a graph. The goal is to make sure that the researchers don't see anything that could prime their minds towards a particular interpretation, such as a preliminary trend or hint of a discovery.

"If the hypotheses are mutually exclusive, you are going to learn something."

Before unblinding data, scientists on Perlmutter's team must circulate a memo explaining their hypotheses and how they plan to test and differentiate between them. "Everybody can decide ahead of time whether that feels fair — that they haven't treated any of the alternatives differently than the others," he says. Last year, Perlmutter and psychologist Robert MacCoun of Stanford University in California argued in a *Nature* Comment[6] that this approach could reduce researcher bias in many fields.

Of course, there are situations in which multiple hypotheses aren't helpful — or even feasible. If researchers stumble on a mysterious finding, they might struggle to come up with even a single plausible explanation. And even if they can cobble together a few, there is no guarantee that the correct hypothesis is among them. This is why hypotheses must remain 'working', so that they can be refined in light of new information.

Other situations present the opposite challenge: too many hypotheses. Freya Blekman, an experimental physicist at the Dutch-speaking Free University of Brussels, searches for elementary particles at facilities such as the Large Hadron Collider at CERN, Europe's particle-physics lab near Geneva, Switerland. In her field, theorists have already posited countless possibilities, and her task is to work out which ones the evidence supports.

Because these models are often mutually exclusive, she typically evaluates them one at a time using *P* values — albeit held to an exceptionally high standard of significance. In fields such as psychology and medicine, there is a growing movement to abandon this technique because it can tempt researchers to seek out analytical approaches that produce significant results. But Blekman says that the physics community has largely eliminated this problem through blinding and by creating a culture so steeped in the ethos of multiple hypotheses that finding nothing is as important as finding something. "In our field, a null result is a valuable result," she says.

Indeed, the method of multiple hypotheses doesn't always have to be practised at the individual level, and can take place across entire

PET IDEAS

Don't play favourites

To apply the multiple-working-hypotheses method, try these tips:

- Devise a list of possible hypotheses before collecting or looking at new data.
- Talk to colleagues and try to challenge your assumptions by creating at least one outrageous hypothesis.
- To learn most efficiently, develop hypotheses that are as distinct from each other as possible.
- Use analytical techniques that block you from developing preliminary ideas about what your data are telling you. This could include analysing samples out of order, blinding your data or using different statistical tests.
- Before looking at your data, try to articulate all possible outcomes, and how you would test and differentiate each one.
- Keep in mind that a null result is not a failure but rather an additional piece of information. J.R.

fields. Different groups can advance various hypotheses, as long as they remain open-minded, and the peer-review process can also help to promote the practice. "I think we have a duty as editors and reviewers to bring up alternatives," says Branch, "and to require authors that come up with a new hypothesis to also include alternatives when they bring it up the first time around".

Regardless of how they apply the method, many researchers say that they stumbled across the idea of multiple hypotheses by accident, as graduate students or later. Branch had never heard of the concept until a few years ago, but was so struck by it that he wrote an article last year arguing that researchers should not seek a single, universal explanation for how fisheries affect marine food webs, but should consider how different models might apply in various parts of the world[7].

A few researchers say that their advisers encouraged them to read classic philosophy-of-science texts, such as Thomas Kuhn's *Structure of Scientific Revolutions* (Univ. Chicago Press, 1962), or fostered discussions on the practical side of the scientific method at lab meetings. But many scientists can make it through their entire careers without any formal training in how to develop hypotheses.

That's too bad, because learning and applying the multiple-hypothesis method can improve the calibre of scientists' work and empower scientists themselves, says Symes, who published a guide last year on teaching the research process[8]. "It always pains me to see students who define success and failure as whether they support a particular hypothesis," she says. "Failing is not collecting the data you need. Succeeding is being able to differentiate the possibilities." ■

Julia Rosen *is a freelance writer in Portland, Oregon.*

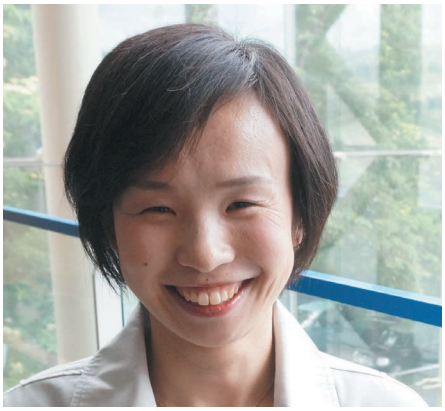
1. Chamberlin, T. C. *Science* **15,** 92–96 (1890).
2. Smaldino, P. E. & McElreath, R. Preprint at https://arxiv.org/abs/1605.09511 (2016).
3. Elliott, L. P. & Brook, B. W. *BioScience* **57,** 608–614 (2007).
4. Pardon, L. G., Brook, B. W., Griffiths, A. D. & Braithwaite, R. W. *J. Animal Ecol.* **72,** 106–115 (2003).
5. Wegener, A. *Petermanns Geogr. Mitt.* **58,** 185–185, 253–256, 305–309 (1912).
6. MacCoun, R. & Perlmutter, S. *Nature* **526,** 187–189 (2015).
7. Branch, T. A. *Fisheries* **40,** 373–375 (2015).
8. Symes, L. B., Serrell, N. & Ayres, M. P. *Bull. Ecol. Soc. Am.* **96,** 352–367 (2015).

CORRECTION

The Careers Feature 'Partners in knowledge' (*Nature* **535,** 581–582; 2016) mistakenly attributed the tradition of depicting unusual events on buffalo hides to the Great Lakes region. It is actually a Great Plains tradition.

TURNING POINT
Planet navigator

Chikako Hirose, an aerospace engineer for the Japan Aerospace Exploration Agency (JAXA), led the team that steered the Akatsuki probe into orbit around Venus on 7 December 2015. She has directed Japan's only successful planetary mission so far, recovering the spacecraft from a failed insertion attempt in 2010.



What led you to become an aerospace engineer?

When I was nine years old, I learned from my schoolteacher that human beings had been to the Moon. I became curious about space. At 15, I sent out letters to many laboratories at NASA, asking for advice on how to get involved in space-related activities. I got lucky — one retired engineer from NASA's Goddard Space Flight Center replied. He told me to study hard in chemistry, physics and mathematics. When I was 19, JAXA announced that 20 students would be selected to attend the 50th International Astronautical Congress in Amsterdam, which I applied for. The opportunity eventually led to an official job offer from JAXA.

Why were you in the control room when Akatsuki failed to enter Venus's orbit in 2010?

I wanted to get involved in deep-space missions. I would go to the Akatsuki project room every day just to see if there was something I could do. Mostly, I just listened. The spacecraft was passing behind Venus when it was set to enter orbit, so we couldn't receive continuous signals. When the predicted time came, we didn't receive anything. One second passed, two, three — after 15 seconds, people were whispering, "What is happening to Akatsuki?" We found out that the main engine hadn't fired as planned, so the spacecraft had gone into safe mode and was tumbling. You could see the disappointment on the faces of the scientists.

How did you end up leading the recovery?

I had done work analysing space debris and estimating its close approach to satellites. This experience made me an expert in trajectory and orbital analysis. We determined, on the basis of the gravity of the Sun and Venus, that Akatsuki would only re-encounter Venus five years later. We tried to preserve the spacecraft as best we could. Its design life was just two and a half years.

What was the key constraint in designing Akatsuki's new trajectory?

The spacecraft's orbit had become very long and elliptical — 370,000 kilometres at its farthest distance from Venus (similar to the distance between Earth and the Moon) and 400 kilometres at its closest. At its farthest point, the spacecraft could take more than ten hours to pass through the planet's shadow. But Akatsuki's solar-charged batteries last for less than two hours. We had to adjust the spacecraft's orbit several times over five years and perform a manoeuvre so as not to exceed Akatsuki's battery life.

How confident were you that the mission would succeed?

I still didn't know whether Akatsuki's engines really worked. Our initial plan was to use the four engines on one side. If they failed, we were prepared to rotate the spacecraft 180 degrees to use the four engines on the other side. We were closely monitoring the velocity of the spacecraft, and saw that the change was exactly as expected. We knew that Akatsuki had entered into orbit around Venus.

How did you celebrate?

In 2010, we had made preparations to celebrate, but failed. In 2015, I had brought a bottle of champagne with me, but didn't tell any of my colleagues until after the operation was complete. We opened the bottle and drank it together.

Are you still involved with Akatsuki?

Yes. I am still responsible for controlling Akatsuki's orientation with respect to Venus, which changes almost every hour when the craft is closest to the planet. I also have to ensure that the spacecraft is oriented correctly for downlinking its observation data to Earth. We expect Akatsuki to survive another five years before crashing into Venus. ■

INTERVIEW BY SMRITI MALLAPATY

This interview has been edited for length and clarity.

fields. Different groups can advance various hypotheses, as long as they remain open-minded, and the peer-review process can also help to promote the practice. "I think we have a duty as editors and reviewers to bring up alternatives," says Branch, "and to require authors that come up with a new hypothesis to also include alternatives when they bring it up the first time around".

Regardless of how they apply the method, many researchers say that they stumbled across the idea of multiple hypotheses by accident, as graduate students or later. Branch had never heard of the concept until a few years ago, but was so struck by it that he wrote an article last year arguing that researchers should not seek a single, universal explanation for how fisheries affect marine food webs, but should consider how different models might apply in various parts of the world[7].

A few researchers say that their advisers encouraged them to read classic philosophy-of-science texts, such as Thomas Kuhn's *Structure of Scientific Revolutions* (Univ. Chicago Press, 1962), or fostered discussions on the practical side of the scientific method at lab meetings. But many scientists can make it through their entire careers without any formal training in how to develop hypotheses.

That's too bad, because learning and applying the multiple-hypothesis method can improve the calibre of scientists' work and empower scientists themselves, says Symes, who published a guide last year on teaching the research process[8]. "It always pains me to see students who define success and failure as whether they support a particular hypothesis," she says. "Failing is not collecting the data you need. Succeeding is being able to differentiate the possibilities." ■

Julia Rosen *is a freelance writer in Portland, Oregon.*

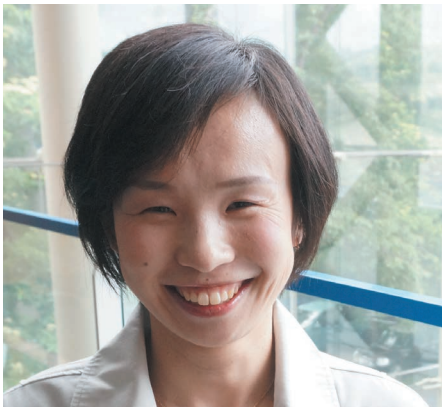
1. Chamberlin, T. C. *Science* **15,** 92–96 (1890).
2. Smaldino, P. E. & McElreath, R. Preprint at https://arxiv.org/abs/1605.09511 (2016).
3. Elliott, L. P. & Brook, B. W. *BioScience* **57,** 608–614 (2007).
4. Pardon, L. G., Brook, B. W., Griffiths, A. D. & Braithwaite, R. W. *J. Animal Ecol.* **72,** 106–115 (2003).
5. Wegener, A. *Petermanns Geogr. Mitt.* **58,** 185–185, 253–256, 305–309 (1912).
6. MacCoun, R. & Perlmutter, S. *Nature* **526,** 187–189 (2015).
7. Branch, T. A. *Fisheries* **40,** 373–375 (2015).
8. Symes, L. B., Serrell, N. & Ayres, M. P. *Bull. Ecol. Soc. Am.* **96,** 352–367 (2015).

CORRECTION

The Careers Feature 'Partners in knowledge' (*Nature* **535,** 581–582; 2016) mistakenly attributed the tradition of depicting unusual events on buffalo hides to the Great Lakes region. It is actually a Great Plains tradition.

TURNING POINT
Planet navigator

Chikako Hirose, an aerospace engineer for the Japan Aerospace Exploration Agency (JAXA), led the team that steered the Akatsuki probe into orbit around Venus on 7 December 2015. She has directed Japan's only successful planetary mission so far, recovering the spacecraft from a failed insertion attempt in 2010.



What led you to become an aerospace engineer?

When I was nine years old, I learned from my schoolteacher that human beings had been to the Moon. I became curious about space. At 15, I sent out letters to many laboratories at NASA, asking for advice on how to get involved in space-related activities. I got lucky — one retired engineer from NASA's Goddard Space Flight Center replied. He told me to study hard in chemistry, physics and mathematics. When I was 19, JAXA announced that 20 students would be selected to attend the 50th International Astronautical Congress in Amsterdam, which I applied for. The opportunity eventually led to an official job offer from JAXA.

Why were you in the control room when Akatsuki failed to enter Venus's orbit in 2010?

I wanted to get involved in deep-space missions. I would go to the Akatsuki project room every day just to see if there was something I could do. Mostly, I just listened. The spacecraft was passing behind Venus when it was set to enter orbit, so we couldn't receive continuous signals. When the predicted time came, we didn't receive anything. One second passed, two, three — after 15 seconds, people were whispering, "What is happening to Akatsuki?" We found out that the main engine hadn't fired as planned, so the spacecraft had gone into safe mode and was tumbling. You could see the disappointment on the faces of the scientists.

How did you end up leading the recovery?

I had done work analysing space debris and estimating its close approach to satellites. This experience made me an expert in trajectory and orbital analysis. We determined, on the basis of the gravity of the Sun and Venus, that Akatsuki would only re-encounter Venus five years later. We tried to preserve the spacecraft as best we could. Its design life was just two and a half years.

What was the key constraint in designing Akatsuki's new trajectory?

The spacecraft's orbit had become very long and elliptical — 370,000 kilometres at its farthest distance from Venus (similar to the distance between Earth and the Moon) and 400 kilometres at its closest. At its farthest point, the spacecraft could take more than ten hours to pass through the planet's shadow. But Akatsuki's solar-charged batteries last for less than two hours. We had to adjust the spacecraft's orbit several times over five years and perform a manoeuvre so as not to exceed Akatsuki's battery life.

How confident were you that the mission would succeed?

I still didn't know whether Akatsuki's engines really worked. Our initial plan was to use the four engines on one side. If they failed, we were prepared to rotate the spacecraft 180 degrees to use the four engines on the other side. We were closely monitoring the velocity of the spacecraft, and saw that the change was exactly as expected. We knew that Akatsuki had entered into orbit around Venus.

How did you celebrate?

In 2010, we had made preparations to celebrate, but failed. In 2015, I had brought a bottle of champagne with me, but didn't tell any of my colleagues until after the operation was complete. We opened the bottle and drank it together.

Are you still involved with Akatsuki?

Yes. I am still responsible for controlling Akatsuki's orientation with respect to Venus, which changes almost every hour when the craft is closest to the planet. I also have to ensure that the spacecraft is oriented correctly for downlinking its observation data to Earth. We expect Akatsuki to survive another five years before crashing into Venus. ■

INTERVIEW BY SMRITI MALLAPATY

This interview has been edited for length and clarity.

WALLS OF NIGERIA

Family ties.

BY JEREMY SZAL

ILLUSTRATION BY JACEY

I stare at the twisted remains of Lagos through the visor of my exosuit as I stalk down the hill. Buildings crumble and slide into the sea. Coils of fiery smoke curl up to the sky. So much work, so much craftsmanship. Gone in weeks.

I'm panting as I continue down the hill — with the cooling system broken, I'm swimming in sweat inside this thing. It's gunmetal grey, covering me from sole to scalp and weighing several hundred kilos. If it weren't for the hydraulics built along my spine, moving in it would be impossible. I have to make extra effort to control it now; the suit seems to have a mind of its own. Cancelling my HUD commands, seizing up at random intervals, cutting off my sensory details.

I'm nearing the school I used to attend, years before any of this happened. A few lone palm trees remain, fronds swaying in the sour wind. I remember being in class one stifling Tuesday, me and Tendai trying to sneak out when we first heard we'd captured one of the K'Dasewh. After all these years, we'd finally got an alien.

There are remains of a solider over by the school. An art mural covers the wall, unfinished words scrawled on blasted brick the colour of red earth. Chalk lies strewn on the ground. Even though his armour has been cracked open, it still pulses with blue bioluminescence. The suit had grown into his flesh like a graft, the metal and matte and wires worming through his dark skin like tendrils. I step over empty coconut shells to check his suit's reading to see when he died. Almost three months ago. He'd been wearing his suit for only two months and he's this far gone.

I've been inside mine for two years.

My skin crawls with the memory of being locked into our suits of armour, laced with alien DNA. They'd dissected these aliens, taken the self-healing and enhanced strength in their biotech and transferred it to us. For a while, it worked.

We didn't know that for the biotech to function and repair us, it needs living tissue. You can't get biomass from nothing. So the suit slowly grew inwards into flesh, tunnelling through open wounds and organs damaged from battle. Fusing into the wearer. The quarantine came around too late.

I wonder if I have any flesh left, if the cables have wrapped around my bones like creepers around a tree. If it's started corroding my brain, trying to take complete control of the suit. Tightening its grip by the day. But I have no way of knowing. And that scares me the most.

Over in the distance, there's a biosphere laid out over the ground — where some of the last human settlements still reside. We're not allowed within five klicks of them for risk of infection. They're still getting refugees from Ghana and Cameroon, but most of them have already been placed on off-world colonies and habitable planets outside the Solar System.

My wife and sons are among them. Ben should be six years old now and Emeka eight, maybe nine.

These are just the last few that have lingered behind on Earth to make sure that no one gets left behind. No one except us.

I log into my commander's channel. It takes me three tries to get it right; the suit attempts to cancel it. But I manage it.

"You still out there, son?" The grizzled face of Commander Somadina pops into my bottom-right vision. "I thought you were dead."

I wish I was. I truly do. "I'm still here."

"I wish I could help. But we can't let any of you Stained inside the sphere. We can't let the biotech virus spread, especially not to the new colony."

My jaws lock and my muscles tighten against my armour. "After everything we did?" I spread my arms, armour plates like fetters around my wrists. "We fought for this city with everything we had."

"And look what happened anyway." He shakes his head. "They sent their entire fleet and destroyed it."

No matter what we did — how hard we fought — it wasn't enough. By the time we'd destroyed the last of their ships, our world was broken.

I crane my neck to look at the sky. Somewhere, out in the giant cosmos of space, is my family. "At least let me talk to my wife one last time. Let me send a message."

"Cannot be done. We can't tell you where the colony is. What if they capture and torture you? Besides, your armour will store the location. All other Stained are in the same position."

I want to scream. I want to laugh like a madman. My throat's filled with concrete and every word feels like it's being fishhooked from my gut. Maybe now the suit has started consuming my throat and vocal cords. Soon I won't be able to speak. "So that's it?"

"I'm so sorry." He can't even look at me. "Goodbye, Kohban."

He cuts the connection. Leaving me here, shackles worming deeper and deeper into my body.

The weight of my armour and of the cosmos pressing down on my shoulders, I stagger to the wall and scoop up some chalk. Hands shaking, I scrawl a message to my friends and family, to the people of Nigeria. I do it quickly, before the armour locks up. Telling them that I miss them — that I'm part of this world now. That the K'Dasewh will never have our planet.

And one day, when my people return to a new, clean Earth, this message will greet them. I hope I'm not here when that happens.

My eyes blur. It could be tears, or could be the suit trying to obscure my vision. I don't think I'll ever know for sure. ■

Jeremy Szal's *work has appeared in* Nature, Abyss & Apex, Lightspeed *and others. He lives in Sydney, Australia, and seeks literary representation. Find him at jeremyszal.com or @jeremyszal*

➲ NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm